The Mamba Mentality

This note synthesizes results on structured and selective state space models. It organizes the literature around recurrent, convolutional, transfer-function, and semiseparable representations; controlled and selective dynamics; state-space duality; expressivity and limitations; stability, generalization, and in-context learning. Detailed proofs are collected in the appendix.

Contents

1 Notation and Conventions

Global sequence notation. Unless a cited theorem is explicitly kept in source notation, sequences are indexed by $t = 1, \dots, L$. Inputs are x_t, hidden states are h_t, and outputs are y_t. Continuous-time signals are written $u(t)$ or X_t, depending on whether the result is an ODE statement or a path/CDE statement. The state dimension is N, the input dimension is d_{in}, and the output dimension is d_{out}.

Symbol	Convention
A, B, C, D	Continuous-time LTI SSM parameters; A is the state matrix, B is the input map, C is the readout, and D is the skip map.
$\bar{A}, \bar{B}$	Discretized parameters. A bar always means discretization, not complex conjugation.
$\bar{A}_t, \bar{B}_t, C_t$	Token-dependent selective parameters. A time subscript on an SSM parameter means input dependence.
Δ_t	Selective step size, usually generated from x_t and passed through softplus.
b_t	Affine scan input, usually $b_t = \bar{B}_t x_t$.
K_j	Convolution kernel coefficient, $K_j = C\bar{A}^j\bar{B}$.
M	Causal lower triangular sequence matrix induced by a recurrent layer.
L	Default sequence length. Some cited generalization and ICL theorems use T or N for sample count, horizon, or context length; those local meanings are stated where they occur.
Z_t^X	Linear CDE state driven by an input path X. It is not the same symbol as the discrete SSM state h_t.
ω^X, ξ^X	CDE gates in the selective-state-space theory of Cirone et al. [4].
$\mathrm{Sig}(X)$	Path signature. Words are written $I = (i_1, \ldots, i_m)$ and $A_I = A_{i_m} \cdots A_{i_1}$.
$\mathcal{A} = (Q, \Sigma, q_{\text{init}}, \delta)$	Deterministic finite automaton. Its one-hot state vector is denoted $\mathbf{q}_t$, not x_t.

Linear time-invariant SSM.

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \qquad y_t = Ch_t + Dx_t.$$

Selective SSM.

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \qquad y_t = C_t h_t + Dx_t.$$

Reading convention. Architecture names enter through assumptions on the recurrence or sequence operator. Long derivations and notation translations are placed in the appendix. When a result is best read in its original letters, the theorem statement says so explicitly; otherwise the notation above is the default.

Attribution convention. Headings with a cited theorem, proposition, lemma, corollary, or section are source results stated in the notation of this note. Headings marked "reformulating" are algebraic restatements of source formulas. Results in a subsection explicitly marked as derived consequences are proved from earlier displayed recurrences in this note. Bare definitions are local conventions or source definitions, as indicated in their headings.

Result Map

Location	Mathematical role
Section 2	Recurrent, convolutional, and transfer-function forms of an SSM layer.
Section 3	HiPPO memory, LSSL/S4 structured parameterization, diagonal SSMs, S5, and H3-style recall.
Section 4	Input-dependent recurrence, Linear CDE selectivity, path signatures, and structured selective-SSM variants.
Section 5	Hidden attention, structured semiseparable matrices, SSD, and the boundary with softmax attention.
Section 6	Formal languages, polynomial interactions, finite automata, negative/complex eigenvalues, and state tracking.
Section 7	Copying lower bounds, uniform circuit containment, metric automata, parity obstructions, and diagonal group-tracking limits.
Section 8	Stable approximation, PAC and covering bounds, Lyapunov perturbation bounds, and simplified Mamba training dynamics.
Section 9	Gradient-descent constructions, Markov-chain ICL, online-gradient behavior, outlier robustness, and low-dimensional target learning.

Model-Class Index

Model object	Where it appears
LTI SSMs, LSSL, S4	Sections 2–3.
Diagonal SSMs, DSS, S4D, DLR	Sections 3, 6, and 7.
S5 and MIMO diagonal SSMs	Section 3.
S6, Mamba, Mamba-2	Sections 4, 5, 8, and 9.
Linear CDE selective SSMs	Section 4.
SSD and semiseparable attention	Section 5.
Selective SSM variants with complex or MIMO structure	Section 4.
Finite-precision LRNNs and SSMs	Sections 6 and 7.

2 Model Objects

2.1 State Space Layers

Definition 2.1 (Continuous LTI SSM, local convention)**.** A continuous-time linear time-invariant state space model is

$$\dot{h}(t) = Ah(t) + Bu(t), \qquad y(t) = Ch(t) + Du(t),$$

with $h(t) \in \mathbb{C}^N$. For a step size $\Delta > 0$, zero-order-hold discretization gives

$$\bar{A} = \exp(\Delta A), \qquad \bar{B} = \int_0^{\Delta} \exp(\tau A) B \, d\tau,$$

and hence

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \qquad y_t = Ch_t + Dx_t.$$

See Appendix A.8. This is the baseline object behind LSSL and S4 [14, 13].

Proposition 2.2 (Reformulated recurrence, convolution, transfer)**.** *For $h_0 = 0$,*

$$y_t = Dx_t + \sum_{j=0}^{t-1} K_j x_{t-j}, \qquad K_j = C\bar{A}^j\bar{B}.$$

Equivalently, the LTI layer is a causal Toeplitz operator with transfer function

$$H(z) = C(zI - \bar{A})^{-1}\bar{B} + D$$

up to the chosen z-normalization convention.

Proof. Appendix A.9.

The LSSL paper makes the continuous, recurrent, and convolutional views explicit [14]. Appendix A.9 records the indexing.

Proposition 2.3 (Transfer coordinates and state-free recovery, [32])**.** *Let*

$$H(z) = h_0 + C(zI - A)^{-1}B$$

be the transfer function of a discrete LTI SSM. If

$$\widehat{A} = SAS^{-1}, \qquad \widehat{B} = SB, \qquad \widehat{C} = CS^{-1}$$

for an invertible S, then the transfer function is unchanged. If

$$H_\ell(z) = \sum_{t=0}^{\ell-1} h_t z^{-t}, \qquad \mathbb{T}_m = \{e^{2\pi ik/m} : 0 \le k < m\},$$

and $m \ge \ell$, then for $0 \le t < \ell$,

$$h_t = \mathsf{iFFT}_m\big((H_\ell(z))_{z\in\mathbb{T}_m}\big)_t.$$

For $h_t = CA^{t-1}B$ when $t > 0$, the transfer-function tail satisfies

$$\sum_{t=\ell+1}^{\infty} CA^{t-1}Bz^{-t} = CA^\ell z^{-\ell}(zI - A)^{-1}B$$

where the resolvent series converges. Finally, a simple-pole decomposition

$$H(z) = h_0 + \sum_{i=1}^{n} \frac{r_i}{z - \lambda_i}$$

has diagonal impulse response $h_t = \sum_i r_i\lambda_i^{t-1}$ for $t > 0$.

Proof. Appendix A.10.

Thus the transfer-function view treats the rational function as the coordinate-invariant object; diagonal SSMs arise by partial fractions, and state-free inference evaluates the truncated transfer function at roots of unity.

2.2 Structured State Matrices

S4. S4 uses normal-plus-low-rank or diagonal-plus-low-rank structure,

$$A = \Lambda - PQ^*,$$

so that resolvent and kernel computations reduce to structured Cauchy-type products [13].

Diagonal SSMs. For $A = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$, the kernel $K_j = C\,\mathrm{diag}(\lambda_1^j, \dots, \lambda_N^j)\bar{B}$ is Vandermonde-type. This is the algebraic setting of DSS and S4D [16, 12]: eigenvalue placement determines the long-range kernel geometry. S5 [38] keeps diagonal dynamics and scan-based recurrence, but uses MIMO input and output maps.

2.3 Selective State Spaces

Definition 2.4 (Selective SSM, local convention)**.** A selective SSM lets the discrete transition, input, and readout maps depend on the current token. At time t it has the form

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \qquad y_t = C_t h_t + D x_t.$$

In Mamba, Δ_t, B_t, and C_t are generated from x_t, while the underlying state matrix A is learned and structured [10].

There is generally no fixed convolution kernel. The scan object is the affine map

$$h \mapsto \bar{A}_t h + b_t, \qquad b_t = \bar{B}_t x_t,$$

with composition

$$(\bar{A}, b) \circ (\bar{A}', b') = (\bar{A}\bar{A}',\ \bar{A}b' + b).$$

Appendix A.11 checks associativity.

2.4 Sequence Operators

Unrolling the selective recurrence gives a causal lower triangular operator $y = Mx$ with blocks

$$M_{ij} = C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j, \qquad i \geq j.$$

For LTI SSMs, M is Toeplitz. For selective SSMs, M is data-dependent. In the SSD/Mamba-2 line, M is studied through semiseparable structure and structured masked attention [1, 7]. See Appendices A.13 and A.14.

3 Linear State Spaces as Memory and Compression

3.1 HiPPO

Definition 3.1 (HiPPO projection, [11])**.** Let $\mu^{(t)}$ be a measure supported on $(-\infty, t]$ and let G be an N-dimensional polynomial subspace. The HiPPO state at time t is the coefficient vector, in an orthogonal polynomial basis of G, of the $L^2(\mu^{(t)})$ projection of the history $u_{\leq t}$ onto G.

Theorem 3.2 (HiPPO-LegS dynamics, Theorem 2 of [11])**.** *For the scaled Legendre measure $\mu^{(t)} = \frac{1}{t}\mathbf{1}_{[0,t]}$, the coefficient state satisfies*

$$\dot{h}(t) = -\frac{1}{t} A h(t) + \frac{1}{t} B u(t),$$

where

$$A_{nk} = \begin{cases} (2n+1)^{1/2}(2k+1)^{1/2}, & n > k, \\ n+1, & n = k, \\ 0, & n < k, \end{cases} \qquad B_n = (2n+1)^{1/2}.$$

The Euler discretization displayed in the source is

$$h_{k+1} = \left(I - \frac{A}{k} \right) h_k + \frac{1}{k} B u_k.$$

Proof. Appendix A.1.1.

Proposition 3.3 (HiPPO-LegS timescale equivariance and fast step, Propositions 3–4 of [11]). *Let* $\mathrm{HiPPO}_N(u)(t)$ *denote the degree-*$(N-1)$ *LegS coefficient state of* u *at time* t*. If* $\alpha > 0$ *and* $v(t) = u(\alpha t)$*, then*

$$\mathrm{HiPPO}_N(v)(t) = \mathrm{HiPPO}_N(u)(\alpha t).$$

Moreover, under any generalized bilinear-transform discretization, one HiPPO-LegS recurrence update can be computed in $O(N)$ *operations.*

Proof. Appendix A.1.2.

Proposition 3.4 (HiPPO-LegS approximation error, Proposition 6 of [11]). *Let* $g^{(t)}$ *be the degree-*$(N-1)$ *HiPPO-LegS projection of* $u_{\le t}$*. If* u *is* L_u*-Lipschitz, then*

$$\|u_{\le t} - g^{(t)}\| = O(tL_u/\sqrt{N}).$$

If u *has order-*k *bounded derivatives, then*

$$\|u_{\le t} - g^{(t)}\| = O(t^k N^{-k+1/2}).$$

Proof. Appendix A.1.3.

Theorem 3.5 (FouT finite-window approximation, [15]). *As* $N \to \infty$*, the FouT SSM is a time-invariant orthogonal SSM for the finite-window measure on* $[0,1]$ *with truncated Fourier basis. Let* $K : [0,1] \to \mathbb{R}$ *be differentiable, and let* $\widehat{K}_N$ *be its representation by the FouT system with state size* N*. If* K *is* L_K*-Lipschitz, then for every* $\varepsilon > 0$*,*

$$N \ge \left(\frac{L_K}{\pi\varepsilon}\right)^2 + 2 \quad \Longrightarrow \quad \|K - \widehat{K}_N\|_2 \le \varepsilon.$$

If K *has* k *derivatives bounded by* L_K*, then it suffices to take*

$$N \ge \left(\frac{L_K}{\pi^k\varepsilon}\right)^{2/(2k-1)} + 2.$$

Proof. Appendix A.1.4.

3.2 LSSL

Proposition 3.6 (Recurrent/convolutional LSSL form, Section 3.1 of [14]). *For a fixed discretized SSM*

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \qquad y_t = Ch_t + Dx_t,$$

with $h_0 = 0$*, the same layer is the causal convolution*

$$y_t = Dx_t + \sum_{j=0}^{t-1} K_j x_{t-j}, \qquad K_j = C\bar{A}^j\bar{B}.$$

Equivalently, the length-L *kernel is the Krylov sequence*

$$\mathcal{K}_L(\bar{A}, \bar{B}, C) = (C\bar{B},\ C\bar{A}\bar{B}, \ldots, C\bar{A}^{L-1}\bar{B}).$$

Proof. Appendix A.1.5.

Lemma 3.7 (Gates as discretization, Lemma 1 of [14]). *The scalar gated recurrence*

$$h_t = (1 - \sigma(z))h_{t-1} + \sigma(z)x_t$$

is the backward-Euler discretization of $\dot{h} = -h + x$ *with step size* $\Delta t = \exp(z)$*.*

Proof. Appendix A.1.6.

Theorem 3.8 (Fast Krylov computation, Theorem 2 of [14]). *For a* k*-quasiseparable matrix* A *with constant* k*, the Krylov function* $\mathcal{K}_L(A, B, C)$ *can be computed in quasi-linear time and space* $\widetilde{O}(N + L)$ *and logarithmic depth in the exact-arithmetic structured-matrix model.*

Proof. Appendix A.1.7.

3.3 S4

Lemma 3.9 (Conjugation invariance, Lemma 3.1 of [13]). *For any invertible V, the SSMs*

$$(A, B, C) \qquad \text{and} \qquad (V^{-1}AV,\ V^{-1}B,\ CV)$$

define the same input-output map.

Proof. Appendix A.1.8.

Lemma 3.10 (Ill-conditioned HiPPO diagonalization, Lemma 3.2 of [13]). *The HiPPO matrix studied in S4 is diagonalized by $V_{ij} = \binom{i+j}{i-j}$. In particular, V contains entries of size approximately $2^{4N/3}$.*

Proof. Appendix A.1.9.

Theorem 3.11 (HiPPO matrices are NPLR, Theorem 1 of [13]). *The HiPPO matrices used in S4, including LegS, LegT, and LagT variants, admit a normal-plus-low-rank representation*

$$A = V\Lambda V^* - PQ^\top,$$

with V unitary, Λ diagonal, and low rank $r = 1$ or $r = 2$. After the unitary change of basis, this becomes diagonal-plus-low-rank.

Proof. Appendix A.1.10.

Theorem 3.12 (S4 recurrent and convolutional complexity, Theorems 2–3 of [13]). *For a constant-rank DPLR state matrix and bilinear discretization, one SSM recurrent step can be computed in $O(N)$ operations. The length-L convolution kernel reduces to four Cauchy matrix-vector products and has complexity $\widetilde{O}(N + L)$ with $O(N + L)$ space under the stated Cauchy primitive.*

Proof. Appendix A.1.11.

3.4 Diagonal State Spaces

Proposition 3.13 (DSS finite-horizon diagonalization, Proposition 1 of [16]). *Let K be the length-L zero-order-hold kernel of an SSM (A, B, C) with $A \in \mathbb{C}^{N\times N}$ diagonalizable, eigenvalues $\lambda_i \neq 0$, and $e^{L\lambda_i\Delta} \neq 1$. If $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$ and $P_{ik} = \lambda_i k\Delta$, then there exist row vectors $\widetilde{w}, w$ such that*

$$K = \bar{K}_{\Delta,L}(\Lambda, \mathbf{1}, \widetilde{w}) = \widetilde{w}\Lambda^{-1}(e^{\Lambda\Delta} - I)\exp(P),$$

and also

$$K = \bar{K}_{\Delta,L}\big(\Lambda, \big((e^{L\lambda_i\Delta} - 1)^{-1}\big)_i, w\big) = w\Lambda^{-1}\operatorname{rowsoftmax}(P).$$

Proof. Appendix A.1.12.

This is a finite-horizon algebraic kernel identity; it does not bound the norms or conditioning of the resulting diagonal parameters.

Proposition 3.14 (S4D Vandermonde kernel, Section 3.2 of [12]). *For a diagonal discretized SSM,*

$$\bar{K}_\ell = \sum_{n=1}^{N} C_n \bar{A}_n^\ell \bar{B}_n,$$

or, in vector form,

$$\bar{K} = (\bar{B}^\top \circ C)\mathcal{V}_L(\bar{A}).$$

Thus diagonal SSM kernel computation is a Vandermonde multiplication problem.

Proof. Appendix A.1.13.

Theorem 3.15 (S4D-LegS limit, Theorem 3 of [12]). *Let $A = A^{(N)} - PP^\top$ be the HiPPO-LegS decomposition used by S4D, with $A^{(N)}$ normal. As $N \to \infty$, the SSM basis generated by $(A^{(N)}, B/2)$ converges to the original HiPPO-LegS basis generated by (A, B). Since $A^{(N)}$ is unitarily diagonalizable, this gives the diagonal HiPPO initialization its limiting interpretation.*

Proof. Appendix A.1.14.

Proposition 3.16 (Diagonal linear RNN equivalence, Proposition 1 of [17]). *For a scalar linear RNN*

$$h_t = Ah_{t-1} + Bx_t, \qquad y_t = Ch_t,$$

if $A = V\,\mathrm{diag}(\lambda_1, \ldots, \lambda_N)V^{-1}$ is diagonalizable over $\mathbb{C}$, then there is a diagonal linear RNN with kernel

$$K_j = \langle w, \Lambda^j \rangle$$

that computes the same input-output map.

Proof. Appendix A.1.15.

Claim 3.17 (Real-positive DLR shift lower bound, Claim 1 of [17]). *For a real-positive DLR with $\lambda_i \in [0, 1]$, representing a one-hot shift kernel at distance N requires coefficients satisfying*

$$\|w\|_\infty \geq 2^{2N - O(\log N)}.$$

Proof. Appendix A.1.16.

Definition 3.18 (Finite-horizon diagonal SSM map, Ran-Milo et al. [34]). For $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, let $A \in \mathbb{K}^{N\times N}$ be diagonal and let $B, C \in \mathbb{K}^N$. The induced SISO LTI map has impulse response

$$k_\ell = C^\top A^\ell B = \sum_{n=1}^{N} C_n B_n A_n^\ell, \qquad \ell \geq 0,$$

and is denoted $\phi_{N,\mathbb{K}}(A, B, C)$ when the field matters.

Proposition 3.19 (Finite-horizon universality, Proposition `result:uni` of Ran-Milo et al. [34]). *Let $\phi : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ be an arbitrary LTI map and let $t \in \mathbb{N}$. For both $\mathbb{K} = \mathbb{R}$ and $\mathbb{K} = \mathbb{C}$, if $N \geq t$, then there are diagonal parameters (A, B, C) over $\mathbb{K}$ such that $\phi_{N,\mathbb{K}}(A, B, C)$ agrees with ϕ up to time t.*

Proof. Appendix A.1.17.

Proposition 3.20 (Complex contains real, Proposition `result:c4r` of Ran-Milo et al. [34]). *Given any real diagonal SSM $\phi_{N_\mathbb{R},\mathbb{R}}(A_\mathbb{R}, B_\mathbb{R}, C_\mathbb{R})$ and any $N_\mathbb{C} \geq N_\mathbb{R}$, there are complex diagonal parameters $(A_\mathbb{C}, B_\mathbb{C}, C_\mathbb{C})$ such that*

$$\phi_{N_\mathbb{C},\mathbb{C}}(A_\mathbb{C}, B_\mathbb{C}, C_\mathbb{C}) = \phi_{N_\mathbb{R},\mathbb{R}}(A_\mathbb{R}, B_\mathbb{R}, C_\mathbb{R}).$$

Proof. Appendix A.1.18.

Theorem 3.21 (Real dimension lower bound for one complex oscillator, Theorem `result:r4c` of Ran-Milo et al. [34]). *Let $t \in \mathbb{N}$ and $\epsilon \geq 0$. Consider a one-dimensional complex diagonal SSM with scalar parameters $(A_\mathbb{C}, B_\mathbb{C}, C_\mathbb{C})$ satisfying*

$$|\sin(\arg A_\mathbb{C})| \geq 0.2, \qquad |A_\mathbb{C}| \geq 0.5^{1/t}, \qquad |B_\mathbb{C} C_\mathbb{C}| \geq 1.$$

If a real diagonal SSM $\phi_{N_\mathbb{R},\mathbb{R}}(A_\mathbb{R}, B_\mathbb{R}, C_\mathbb{R})$ ϵ-approximates $\phi_{1,\mathbb{C}}(A_\mathbb{C}, B_\mathbb{C}, C_\mathbb{C})$ up to time t, then

$$N_\mathbb{R} \geq \lfloor t/9 \rfloor - 1 - 4\epsilon.$$

Proof. Appendix A.1.19.

Theorem 3.22 (Forward-difference lower bound, Theorem `result:r4general` of Ran-Milo et al. [34]). *Let $\phi_{N_\mathbb{R},\mathbb{R}}(A_\mathbb{R}, B_\mathbb{R}, C_\mathbb{R})$ ϵ-approximate an LTI map ϕ up to time t. Let $\phi(\mathbf{I})|_{\mathrm{odd}}$ and $\phi(\mathbf{I})|_{\mathrm{even}}$ denote the odd and even subsequences of the impulse response, and let $(\cdot)_m^{(d)}$ denote the mth entry of the dth forward difference. Then*

$$N_\mathbb{R} \|C_\mathbb{R}^\top \odot B_\mathbb{R}\|_\infty \geq \max_{\substack{d,m\in\mathbb{N},\ d+m\leq\lfloor t/2\rfloor \\ \sigma\in\{\mathrm{odd},\mathrm{even}\}}} 2^{d+2\min\{d,m\}} \left(2^{-d}\left|\left(\phi(\mathbf{I})|_\sigma\right)_m^{(d)}\right| - \epsilon\right).$$

Proof. Appendix A.1.20.

Proposition 3.23 (Complex bounded-parameter construction, Proposition `result:c4general` of Ran-Milo et al. [34]). *Let ϕ be an LTI map and let $t \in \mathbb{N}$. For any $N_{\mathbb{C}} \geq t$, there are complex diagonal parameters $(A_{\mathbb{C}}, B_{\mathbb{C}}, C_{\mathbb{C}})$ that match ϕ up to time t and satisfy*

$$\|B_{\mathbb{C}}\|_2 \leq 2\|\phi(\mathbf{I})_{:t}\|_2, \qquad \|C_{\mathbb{C}}^\top\|_2 \leq 1.$$

Proof. Appendix A.1.21.

3.5 MIMO and Scan

Proposition 3.24 (S5 scan and complexity, Proposition 1 of [38]). *S5 writes a diagonalized MIMO SSM as*

$$\dot{h}(t) = \Lambda h(t) + \widetilde{B}x(t), \qquad y(t) = \widetilde{C}h(t) + Dx(t),$$

and evaluates the discretized recurrence by associative scan. With latent size $P = O(H)$, the paper's accounting gives the same order runtime and memory as S4 for both recurrent and offline sequence evaluation.

Proof. Appendix A.1.22.

Proposition 3.25 (Constrained S4/S5 relationship, Proposition 2 of [38]). *If H S4 systems share the same state matrix and timescale, then the corresponding constrained S5 system with $P = N$ produces outputs that are a linear projection of the H constrained S4 latent states. These outputs are not the ordinary block-diagonal S4 outputs unless the output projection is also changed accordingly.*

Proof. Appendix A.1.23.

3.6 Early Language-Modeling Bridge

Proposition 3.26 (H3 associative recall construction, [8]). *For the four-key/four-value associative-recall language Λ studied in H3, there is an explicit H3 construction with embedding dimension $d = 8$, SSM state dimension $m = 2$, and $H = 4$ heads that solves the task. The construction uses one shift SSM, one diagonal cumulative-sum SSM, multiplicative query-key-value interactions, and a fixed final decoder from the binary value encoding.*

Proof. Appendix A.1.24.

4 Selective State Space Models

4.1 S6 and Selective Scan

Definition 4.1 (S6 selective recurrence; Gu and Dao [10]). In the Mamba S6 layer, the continuous state matrix A is learned and structured, while Δ_t, B_t, and C_t are generated from the token x_t [10]. Zero-order-hold discretization gives

$$\bar{A}_t = \exp(\Delta_t A), \qquad \bar{B}_t = \int_0^{\Delta_t} \exp(\tau A) B_t \, d\tau,$$

and hence

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \qquad y_t = C_t h_t.$$

Proposition 4.2 (Scalar gate, Theorem 1 of Gu and Dao [10]). *Let $N = 1$, $A = -1$, $B = 1$, and $\Delta_t = \mathrm{softplus}(a^\top x_t)$. Under zero-order-hold discretization,*

$$\bar{A}_t = \exp(-\Delta_t), \qquad \bar{B}_t = 1 - \bar{A}_t.$$

Writing $g_t = \sigma(a^\top x_t)$ gives

$$\bar{A}_t = 1 - g_t, \qquad \bar{B}_t = g_t, \qquad h_t = (1 - g_t)h_{t-1} + g_t x_t.$$

Proof. Appendix A.12.

Proposition 4.3 (Derived affine selective scan)**.** *For affine steps $F_{(\bar{A},b)}(h) = \bar{A}h + b$, define*

$$(\bar{A}, b) \circ (\bar{A}', b') = (\bar{A}\bar{A}',\ \bar{A}b' + b).$$

Then $\circ$ is associative. Consequently the prefix product

$$(\bar{A}_t, b_t) \circ \cdots \circ (\bar{A}_1, b_1), \qquad b_t = \bar{B}_t x_t,$$

computes the map from h_0 to h_t for the time-varying selective recurrence.

Proof. Appendix A.11.

4.2 Input Selectivity as Basis Selection and Memory Freezing

Definition 4.4 (Mamba basis function, Huang et al. [21])**.** Consider a scalar continuous-time S6 channel with $A_n = -\lambda_n$, $\lambda_n > 0$, fixed input matrix coefficient B_n, and input dependence only through the step-size gate Δ_n. Its nth hidden coordinate may be written as

$$h_n^{\mathrm{M}}(t) = \int_0^t g_n^{\mathrm{M}}(s; t, x) x(s)\, ds, \qquad g_n^{\mathrm{M}}(s; t, x) = B_n \exp\left(-\lambda_n \int_s^t \Delta_n(x_r)\, dr\right).$$

The corresponding S4D basis is the special case $\Delta_n \equiv 1$:

$$g_n^{\mathrm{S4D}}(s; t) = B_n e^{-\lambda_n (t-s)}.$$

Theorem 4.5 (Haar wavelets from Mamba bases, Huang et al. [21])**.** *Let $\psi_{j,k} : [0,1] \to \mathbb{R}$ be the Haar wavelet*

$$\psi_{j,k}(s) = 2^{j/2}\Big(\mathbf{1}_{[2^{-j}k,\, 2^{-j}(k+1/2))}(s) - \mathbf{1}_{[2^{-j}(k+1/2),\, 2^{-j}(k+1)]}(s)\Big).$$

Let $\widetilde{x}_s = (x_s, s)$ be the input augmented with time. For every $\epsilon > 0$, there are three Mamba basis functions $g_{j,k}^{\mathrm{M}_1}$, $g_{j,k}^{\mathrm{M}_2}$, and $g_{j,k}^{\mathrm{M}_3}$ of Definition 4.4, evaluated at $t = 1$ on $\widetilde{x}$, such that

$$\sup_{s\in[0,1]} \left|\psi_{j,k}(s) - \left(g_{j,k}^{\mathrm{M}_1}(s; 1, \widetilde{x}) - 2g_{j,k}^{\mathrm{M}_2}(s; 1, \widetilde{x}) + g_{j,k}^{\mathrm{M}_3}(s; 1, \widetilde{x})\right)\right| < \epsilon.$$

Proof. Appendix A.2.1.

Corollary 4.6 (Adaptive discontinuous approximation, Huang et al. [21])**.** *Let $\rho : [0,1] \to \mathbb{R}$ be piecewise constant with $m \geq 1$ discontinuities. There exist N Mamba basis functions such that*

$$\left\| \rho - \sum_{n=1}^{N} g_n^{\mathrm{M}} \right\|_{L^2} = O\left(2^{-N/(3m)}\right).$$

In the same comparison made in the source, S4D/Fourier-type bases achieve $O(N^{-1})$ for this approximation problem.

Proof. Appendix A.2.2.

Lemma 4.7 (Sensitivity decay, Huang et al. [21])**.** *Consider a diagonal time-varying recurrence*

$$h_t = \bar{\Lambda}_t h_{t-1} + \bar{B}_t x_t.$$

Assume the maps $x_j \mapsto \bar{\Lambda}_j$ and $x_j \mapsto \bar{B}_j x_j$ are differentiable at the point considered. For $j < t$,

$$\frac{\partial h_t}{\partial x_j} = \left(\prod_{r=j+1}^{t} \bar{\Lambda}_r\right)\left[\frac{\partial}{\partial x_j}(\bar{B}_j x_j) + \frac{\partial \bar{\Lambda}_j}{\partial x_j} \sum_{s=1}^{j-1} \left(\prod_{r=s+1}^{j-1} \bar{\Lambda}_r\right) \bar{B}_s x_s\right].$$

Consequently, for the nth coordinate of an S4D state and an S6/Mamba state, there are prefactors $\widetilde{c}_{\mathrm{S4D}}$ and $\widetilde{c}_{\mathrm{M}}$ independent of t such that

$$\left|\frac{\partial h_t^{\mathrm{S4D},n}}{\partial x_j}\right| = \widetilde{c}_{\mathrm{S4D}}(\lambda_n, B_n, x_{\leq j}) e^{-\lambda_n (t-j-1)}$$

and

$$\left|\frac{\partial h_t^{\mathrm{M},n}}{\partial x_j}\right| = \widetilde{c}_{\mathrm{M}}(\Delta, \lambda_n, B_n, x_{\leq j}) \exp\left(-\lambda_n \sum_{r=j+1}^{t} \Delta(x_r)\right).$$

Proof. Appendix A.2.3.

Lemma 4.8 (Freezing selective time, Huang et al. [21]). *For the discrete S6 recurrence of Lemma 4.7, assume the nth coordinate satisfies*

$$\lim_{t\to\infty} \lambda_n \sum_{r=j+1}^{t} \Delta(x_r) \le c$$

for some $c \ge 0$. Then

$$\liminf_{t\to\infty} \left| \frac{\partial h_t^{\mathrm{M},n}}{\partial x_j} \right| \ge e^{-c} \left| \frac{\partial}{\partial x_j} \big(B_n(x_j)\Delta(x_j)x_j \big) \right|.$$

Thus retaining nonzero sensitivity to x_j along long sequences requires the cumulative selective time $\sum_{r=j+1}^{t} \Delta(x_r)$, or equivalently the product of decay factors, to remain controlled on the retained segment.

Proof. Appendix A.2.4.

Theorem 4.9 (MQAR constructions, Huang et al. [21]). *Consider the multi-query associative-recall task with κ key-value pairs over vocabulary V. The source gives the following one-layer exact constructions.*

1. *A Mamba model without the output gating branch solves MQAR with embedding size $d = O(\kappa + \log|V|)$ and state size $N = \kappa$.*
2. *A Mamba-2 model without the output gating branch solves MQAR with embedding size $d = O(\log\kappa + \log|V|)$ and state size $N = O(\log\kappa)$.*
3. *A Mamba block with an S4D mixer solves MQAR with embedding size $d = O(\kappa \log|V|)$ and state size $N = 1$.*

Proof. Appendix A.2.5.

Lemma 4.10 (Induction head via state-dimension selectivity, Huang et al. [21]). *Let V be a finite vocabulary. Consider the variant selective mixer*

$$h_t = \exp(\Lambda \odot (\mathbf{1}_d \otimes \Delta(\widehat{x}_t))) \odot h_{t-1} + \widehat{x}_t \otimes B_t,$$

where the selective gate acts along the state dimension. There is a one-layer Mamba model with this mixer that solves the induction-head task with embedding size $d = 2|V|$ and state size $N = |V|$.

Proof. Appendix A.2.6.

Remark. *The approximation and task constructions in this subsection are exact representability statements. Several parameter choices use limiting gates such as $\Delta \to 0$ or $\Delta \to \infty$; the statements are not optimization or typical-training guarantees.*

4.3 Linear CDE Formulation

Definition 4.11 (Linear CDE, Definition 3.1 of Cirone et al. [4]). Let

$$\mathcal{X} \subset C_{1,0}([0,1];\mathbb{R}^d)$$

be an input path class. Let

$$\omega : \mathcal{X} \to C_{1,0}([0,1];\mathbb{R}^{d_\omega}), \qquad \xi : \mathcal{X} \to C_{1,0}([0,1];\mathbb{R}^{d_\xi})$$

be continuous gates, and write $\omega^X = \omega(X)$, $\xi^X = \xi(X)$. A Linear CDE with hidden dimension N is

$$dZ_t^X = \sum_{i=1}^{d_\omega} A_i Z_t^X \, d\omega_t^{X,i} + B \, d\xi_t^X, \qquad Z_0^X = Z_0 \in \mathbb{R}^N,$$

where $A_i \in \mathbb{R}^{N\times N}$ and $B \in \mathbb{R}^{N\times d_\xi}$.

Proposition 4.12 (S4 and S6 as Linear CDEs, Section 3 of Cirone et al. [4]). *For $d = 1$, a continuous-time S4 channel*

$$dZ_t = AZ_t\,dt + BX_t\,dt$$

is the Linear CDE of Definition 4.11 with

$$\omega_t^X = t, \qquad \xi_t^X = \int_0^t X_s\,ds.$$

For a Mamba/S6 channel i, let

$$\Delta_i(x) = \mathrm{softplus}(\alpha_i \cdot x + \beta_i) \simeq \delta\,\sigma(\widetilde{\alpha}_i \cdot x + \widetilde{\beta}_i), \qquad \sigma = \mathrm{ReLU},$$

with $\widetilde{\alpha}_i = \alpha_i/\delta$ and $\widetilde{\beta}_i = \beta_i/\delta$. In the $\delta \to 0$ Euler limit,

$$dZ_{i,t}^X = A_i Z_{i,t}^X \sigma(\widetilde{\alpha}_i \cdot X_t + \widetilde{\beta}_i)\,dt + (BX_t)X_t^i \sigma(\widetilde{\alpha}_i \cdot X_t + \widetilde{\beta}_i)\,dt.$$

Thus the channel is a Linear CDE with

$$\omega_{i,t}^X = \int_0^t \sigma(\widetilde{\alpha}_i \cdot X_s + \widetilde{\beta}_i)\,ds, \qquad \xi_{i,t}^X = \int_0^t X_s X_s^i \sigma(\widetilde{\alpha}_i \cdot X_s + \widetilde{\beta}_i)\,ds.$$

Proof. Appendix A.2.7.

Proposition 4.13 (Signature expansion, Cirone et al. [4]). *Let $\mathbb{W}_{d_\omega}$ be the set of words over $\{1, \dots, d_\omega\}$. For $I = (i_1, \dots, i_m)$, write $A_I = A_{i_m} \cdots A_{i_1}$ and $A_\emptyset = I_N$. If B_j is the jth column of B, then the solution of the Linear CDE satisfies*

$$\begin{aligned} Z_t^X = &\sum_{I \in \mathbb{W}_{d_\omega}} A_I Z_0\ \mathrm{Sig}(\omega^X)_{0,t}^I \\ &+ \sum_{j=1}^{d_\xi} \sum_{I \in \mathbb{W}_{d_\omega}} A_I B_j \int_0^t \mathrm{Sig}(\omega^X)_{s,t}^I\,d\xi_s^{X,j}. \end{aligned}$$

Proof. Appendix A.2.8.

Definition 4.14 (Initial-value feature map, Appendix of Cirone et al. [4]). Let an initial-value gate

$$(\cdot)_0 : \mathcal{X} \to \mathbb{R}^{d_0}, \qquad X \mapsto X_0$$

be fixed. The source feature map at time t is

$$T(X)_{0,t} = \left(X_0^i\,\mathrm{Sig}(\omega^X)_{0,t}^I,\ \int_0^t \mathrm{Sig}(\omega^X)_{s,t}^I\,d\xi_s^{X,j} \right)_{i,I,j},$$

with $i \in \{1, \dots, d_0\}$, $j \in \{1, \dots, d_\xi\}$, and $I \in \mathbb{W}_{d_\omega}$.

Proof. Appendix A.2.9.

Proposition 4.15 (RKHS of Linear CDE features; Cirone et al. [4], Appendix). *Let $\mathcal{H}_t^{0,\omega,\xi}$ be the RKHS generated by $X \mapsto T(X)_{0,t}$. A function $F(\cdot, t) : \mathcal{X} \to \mathbb{R}$ belongs to $\mathcal{H}_t^{0,\omega,\xi}$ if and only if*

$$F(X,t) = \sum_{i=1}^{d_0} X_0^i \langle \alpha_i, \mathrm{Sig}(\omega^X)_{0,t} \rangle_{\ell^2} + \sum_{j=1}^{d_\xi} \int_0^t \langle \beta_j, \mathrm{Sig}(\omega^X)_{s,t} \rangle_{\ell^2}\,d\xi_s^{X,j}$$

for some $\alpha_i, \beta_j \in \ell^2(\mathbb{W}_{d_\omega})$. Moreover the RKHS norm is the minimum value of

$$\sum_{i=1}^{d_0} \|\alpha_i\|_{\ell^2(\mathbb{W}_{d_\omega})}^2 + \sum_{j=1}^{d_\xi} \|\beta_j\|_{\ell^2(\mathbb{W}_{d_\omega})}^2$$

over all coefficient families representing $F(\cdot, t)$.

Proof. Appendix A.2.9.

Lemma 4.16 (Kernel formula for T; Cirone et al. [4], Appendix)**.** *For $X, Y \in \mathcal{X}$, define*

$$\mathcal{K}^{X,Y}(s,t) = \langle T(X)_{0,s}, T(Y)_{0,t} \rangle_{\ell^2}.$$

Then

$$\mathcal{K}^{X,Y}(s,t) = \langle X_0, Y_0 \rangle \langle \mathrm{Sig}(\omega^X)_{0,s}, \mathrm{Sig}(\omega^Y)_{0,t} \rangle_{\ell^2} + \int_{\eta=0}^{s} \int_{\tau=0}^{t} \langle \mathrm{Sig}(\omega^X)_{\eta,s}, \mathrm{Sig}(\omega^Y)_{\tau,t} \rangle_{\ell^2} \langle d\xi^X_\eta, d\xi^Y_\tau \rangle.$$

Equivalently, $\mathcal{K}$ is the solution of the two-parameter equation

$$\mathcal{K}^{X,Y}(s,t) = \langle X_0, Y_0 \rangle + \langle \xi^X_s, \xi^Y_t \rangle + \int_{\eta=0}^{s} \int_{\tau=0}^{t} \mathcal{K}^{X,Y}(\eta,\tau) \langle d\omega^X_\eta, d\omega^Y_\tau \rangle.$$

Proof. Appendix A.2.9.

4.4 Closure Theorems

Theorem 4.17 (Dense Linear CDE closure, Theorem 4.1 of Cirone et al. [4])**.** *Let $\mathcal{X} \subset C_{1,0}([0,1]; \mathbb{R}^d)$ be compact. Let ω, ξ be continuous gates satisfying*

$$\omega^{X,1}_t = t, \qquad \omega^{X,2}_t = t^2.$$

Let

$$\Psi : C_{1,0}([0,1]; \mathbb{R}^{d_\omega}) \to \mathbb{R}, \qquad \Phi : C_{1,0}([0,1]; \mathbb{R}^{d_\omega}) \to \mathbb{R}^{d_\xi}$$

be continuous. For every $\epsilon > 0$ there exist N, dense matrices $A_1, \dots, A_{d_\omega}$, B, and a linear readout $C \in \mathbb{R}^{1 \times N}$ such that

$$\sup_{(X,t) \in \mathcal{X} \times [0,1]} \left| C Z^X_t - \Psi(\omega^X_{[0,t]}) - \int_0^t \Phi(\omega^X_{[s,t]}) \cdot d\xi^X_s \right| < \epsilon.$$

Conversely, every output CZ^X_t of a dense Linear CDE is of this path-functional form.

Proof. Appendix A.2.10.

Theorem 4.18 (Full source closure with initial-value gate; Cirone et al. [4], Appendix theorem)**.** *Let $\mathcal{K} \subseteq \mathcal{X}$ be compact and let $(\cdot)_0, \omega, \xi$ be continuous gates with*

$$\omega^{X,1}_t = t, \qquad \omega^{X,2}_t = t^2.$$

For every $\epsilon > 0$ and every

$$F(X,t) = \Psi(\omega^X_{[0,t]}) \cdot X_0 + \int_0^t \Phi(\omega^X_{[s,t]}) \cdot d\xi^X_s,$$

where

$$\Psi \in C^0(C_{1,0}([0,1]; \mathbb{R}^{d_\omega}); \mathbb{R}^{d_0}), \qquad \Phi \in C^0(C_{1,0}([0,1]; \mathbb{R}^{d_\omega}); \mathbb{R}^{d_\xi}),$$

there exist N, dense matrices $A_i \in \mathbb{R}^{N \times N}$, $B \in \mathbb{R}^{N \times d_\xi}$, an initial map $C_0 \in \mathbb{R}^{N \times d_0}$, and a readout $v \in \mathbb{R}^N$ such that the solution of

$$dZ^X_t = \sum_{i=1}^{d_\omega} A_i Z^X_t \, d\omega^{X,i}_t + B \, d\xi^X_t, \qquad Z^X_0 = C_0 X_0,$$

satisfies

$$\sup_{(X,t) \in \mathcal{K} \times [0,1]} |F(X,t) - \langle v, Z^X_t \rangle| \leq \epsilon.$$

Moreover, under the source LeCun initialization

$$[A_i]_{n,n'} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/N), \qquad [C_0]_{n,i}, [B]_{n,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

the probability that such a readout v exists tends to one as $N \to \infty$.

If the matrices A_i are constrained to be diagonal, the coordinates t, t^2 are not required and the corresponding closure is the smaller family

$$F(X,t) = \psi(\omega_t^X) \cdot X_0 + \int_0^t \phi(\omega_t^X - \omega_s^X) \cdot d\xi_s^X,$$

with $\psi \in C^0(\mathbb{R}^{d_\omega}; \mathbb{R}^{d_0})$ and $\phi \in C^0(\mathbb{R}^{d_\omega}; \mathbb{R}^{d_\xi})$. In both the dense and diagonal cases, fixed readouts of fixed Linear CDE parameters lie in the uniform closure of the corresponding displayed family.

Proof. Appendix A.2.11.

Theorem 4.19 (Randomized Linear CDE closure, Theorem 4.2 of Cirone et al. [4])**.** *Under the hypotheses of Theorem 4.17, sample*

$$[A_j]_{n,n'} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/N), \qquad [Z_0]_n, [B]_{n,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

For every functional

$$F(X,t) = \Psi(\omega^X_{[0,t]}) + \int_0^t \Phi(\omega^X_{[s,t]}) \cdot d\xi_s^X$$

of the preceding form and every $\epsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left[\exists C \in \mathbb{R}^{1\times N} : \sup_{(X,t)\in\mathcal{X}\times[0,1]} |F(X,t) - CZ_t^X| \le \epsilon\right] = 1.$$

Proof. Appendix A.2.12.

Theorem 4.20 (Diagonal Linear CDE closure, Theorem 4.3 of Cirone et al. [4])**.** *If $A_1, \dots, A_{d_\omega}$ are constrained to be diagonal, the closure of linear readouts of Linear CDEs is*

$$\left\{(X,t) \mapsto \psi(\omega_t^X) + \int_0^t \phi(\omega_t^X - \omega_s^X) \cdot d\xi_s^X\right\},$$

where

$$\psi : \mathbb{R}^{d_\omega} \to \mathbb{R}, \qquad \phi : \mathbb{R}^{d_\omega} \to \mathbb{R}^{d_\xi}$$

are continuous. In this diagonal theorem the technical coordinates $\omega_t^{X,1} = t$ and $\omega_t^{X,2} = t^2$ are not required.

Proof. Appendix A.2.13.

Corollary 4.21 (Mamba channel closure, Corollary 4.4 of Cirone et al. [4])**.** *In the Mamba channel-separated diagonal setting, the closure reduces to*

$$\left\{(X,t) \mapsto \sum_{i=1}^{d_\omega} \psi_i(\omega_t^{X,i}) + \sum_{i=1}^{d_\omega} \int_0^t \phi_i(\omega_t^{X,i} - \omega_s^{X,i})\, d\xi_s^{X,i}\right\},$$

where $\psi_i, \phi_i : \mathbb{R} \to \mathbb{R}$ are continuous.

Proof. Appendix A.2.14.

Remark (Mamba sign pattern)**.** *In the source Mamba specialization, the diagonal system has*

$$d\omega_t^X = \mathrm{softplus}(WX_t + \lambda)\, dt, \qquad d\xi_t^X = X_t \odot d\omega_t^X, \qquad V = -v \otimes \mathbf{1}_{d_\omega}, \quad v \ge 0.$$

Thus the diagonal homogeneous factor $\exp(\mathrm{diag}(V(\omega_t^X - \omega_s^X)))$ has nonexpansive real modes when ω^X is nondecreasing componentwise. This is the source conditioning observation behind the diagonal Mamba case.

Proposition 4.22 (Input-driven signature recovery by chained diagonal CDEs, Appendix of Cirone et al. [4])**.** *Let $\mathcal{X} \subset C_{1,0}([0,1]; \mathbb{R}^d)$ be compact. We use the source chaining condition in the displayed input-driven specialization: at each stage the gate and injection drivers both contain the needed coordinates of X, and a constant coordinate is adjoined if the source initial-coordinate convention is used. For every word $I \in \mathbb{W}_d$ with $|I| \ge 2$ and every $\epsilon > 0$, there are linear maps $W_k \in \mathbb{R}^{N_k \times 1}$ and diagonal Linear CDE weights such that*

$$dZ_t^{1,X} = \sum_{i=1}^d A_i^{(1)} Z_t^{1,X}\, dX_t^i + B^{(1)}\, dX_t, \qquad Z_0^{1,X} = Z_0^1,$$

and

$$dZ_t^{k+1,X} = \sum_{i=1}^{d+1} A_i^{(k+1)} Z_t^{k+1,X} d \begin{bmatrix} W_k Z^{k,X} \\ X \end{bmatrix}_t^i + B^{(k+1)}\, dX_t, \qquad Z_0^{k+1,X} = Z_0^{k+1}.$$

For some $v \in \mathbb{R}^{N_{|I|-1}}$,

$$\sup_{(X,t)\in\mathcal{X}\times[0,1]} \left| \mathrm{Sig}(X)_{0,t}^I - \langle v, Z_t^{|I|-1,X} \rangle \right| \le \epsilon.$$

Proof. Appendix A.2.15.

Proposition 4.23 (ReLU-gated chained recovery, Appendix of Cirone et al. [4])**.** *Let* $\mathcal{X} \subset C_{1,0}([0,1];\mathbb{R}^d)$ *be compact and suppose a diagonal chained Linear CDE is driven by a nonnegative gate*

$$\omega_t^X = \int_0^t \mathrm{ReLU}(WX_s + b)\, ds.$$

Assume the gate contains paired coordinates so that a fixed linear map L *satisfies*

$$L\omega_t^X = \widetilde{\omega}_t^X, \qquad \widetilde{\omega}_t^X = \int_0^t X_s\, ds,$$

and assume the injection driver contains the same coordinates needed for $d\widetilde{\omega}^X$*. Then for every word* $I \in \mathbb{W}_d$*,* $|I| \ge 2$*, and every* $\epsilon > 0$*, a sufficiently deep chain of diagonal Linear CDEs with linear maps between layers has a linear readout approximating*

$$\mathrm{Sig}(\widetilde{\omega}^X)_{0,t}^I$$

uniformly on $\mathcal{X} \times [0,1]$ *within* ϵ*.*

Proof. Appendix A.2.16.

Proposition 4.24 (Chained diagonal approximation, Proposition 4.5 of Cirone et al. [4])**.** *Let* F *be any functional in the dense class of Theorem 4.17. For every* $\epsilon > 0$*, a sequence of diagonal Linear CDEs chained by linear maps can be chosen so that, eventually in the chain depth* k*, there exists* $C_k \in \mathbb{R}^{1\times N_k}$ *with*

$$\sup_{(X,t)\in\mathcal{X}\times[0,1]} |F(X,t) - C_k Z_t^{k,X}| \le \epsilon.$$

Proof. Appendix A.2.17.

Proposition 4.25 (Path-to-path approximation, Proposition 5.1 of Cirone et al. [4])**.** *Fix a compact* $\mathcal{K} \subseteq \mathcal{X}$ *and continuous maps* $(\cdot)_0, \omega, \xi$ *with* $X_0^1 \equiv 1$ *and* $\omega_t^{X,1} = t$*. For every* $\epsilon > 0$ *and every causal continuous*

$$G : C_{1,0}([0,1];\mathbb{R}^{d_\omega}) \times [0,1] \to \mathbb{R},$$

there exist N*, a feed-forward network* $F_{\mathrm{mlp}} : \mathbb{R}^N \to \mathbb{R}$*, and Linear CDE parameters* C_0, A_i, B*, with* $Z_0^X = C_0 X_0$*, such that*

$$\sup_{(X,t)\in\mathcal{K}\times[0,1]} |F_{\mathrm{mlp}}(Z_t^X) - G(\omega^X, t)| < \epsilon.$$

Proof. Appendix A.2.18.

Remark. *The dense closure theorem uses recurrent mixing through dense* A_i*. The Mamba specialization above is the diagonal, channel-separated specialization; the chained result is the theorem that recovers dense-class signature information through depth.*

4.5 Mamba-3 as an SSM Object

Lahoti et al. [26] modify the Mamba-2/SSD object along three mathematical axes: the discretization rule, the spectrum of the transition, and the input-output rank of the state-space map.

Proposition 4.26 (Exponential-trapezoidal discretization; source proposition of Lahoti et al. [26])**.** *Consider the time-varying linear SSM*

$$\dot{h}(t) = A(t)h(t) + B(t)x(t), \qquad h(t) \in \mathbb{R}^N,$$

with scalar decay $A(t) \in \mathbb{R}$ and grid $\tau_t - \tau_{t-1} = \Delta_t$. Replacing $A(s)$ on $[\tau_{t-1}, \tau_t]$ by $A_t = A(\tau_t)$ and approximating the state-input integral by the generalized trapezoidal rule gives

$$h_t = e^{\Delta_t A_t} h_{t-1} + (1-\lambda_t)\Delta_t e^{\Delta_t A_t} B_{t-1} x_{t-1} + \lambda_t \Delta_t B_t x_t,$$

where $\lambda_t \in [0,1]$ is data-dependent. Equivalently,

$$h_t = \alpha_t h_{t-1} + \beta_t B_{t-1} x_{t-1} + \gamma_t B_t x_t, \quad \alpha_t = e^{\Delta_t A_t}, \quad \beta_t = (1-\lambda_t)\Delta_t e^{\Delta_t A_t}, \quad \gamma_t = \lambda_t \Delta_t.$$

If the adjusted integrand is C^3 with bounded derivatives and $\lambda_t = \frac{1}{2} + O(\Delta_t)$, then the local state-input quadrature error is $O(\Delta_t^3)$. The choice $\lambda_t = 1$ recovers exponential Euler.

Proof. Appendix A.2.19.

Proposition 4.27 (Mamba-3 mask factorization, reformulating Lahoti et al. [26])**.** *Let $v_t = B_t x_t$ and write the exponential-trapezoidal recurrence as*

$$h_0 = \gamma_0 v_0, \qquad h_t = \alpha_t h_{t-1} + \beta_t v_{t-1} + \gamma_t v_t \quad (1 \le t \le T).$$

With empty products equal to 1, define lower-triangular time masks

$$(L_\alpha)_{ts} = \begin{cases} \prod_{r=s+1}^{t} \alpha_r, & s \le t, \\ 0, & s > t, \end{cases} \qquad D_{\beta,\gamma}(t,s) = \begin{cases} \gamma_t, & s = t, \\ \beta_t, & s = t-1, \\ 0, & \text{otherwise.} \end{cases}$$

Then the full time map is

$$(h_0, \dots, h_T)^\top = L_\alpha D_{\beta,\gamma} (v_0, \dots, v_T)^\top.$$

Thus the Mamba-3 structured mask is a one-semiseparable decay mask composed with a two-band state-input mask; in SSD notation it replaces the Mamba-2 diagonal input mask by a bidiagonal one.

Proof. Appendix A.2.20.

Proposition 4.28 (Complex-to-real SSM equivalence; source proposition of Lahoti et al. [26])**.** *Let $m = N/2$ and consider the complex SSM*

$$\dot{z}(t) = \mathrm{diag}(a(t) + i\theta(t))z(t) + (B(t) + i\widehat{B}(t))x(t), \qquad y(t) = \Re\big((C(t) + i\widehat{C}(t))^\top z(t)\big),$$

where $z(t) \in \mathbb{C}^m$, $a(t) \in \mathbb{R}$ is scalar, and $\theta(t), B(t), \widehat{B}(t), C(t), \widehat{C}(t) \in \mathbb{R}^m$. Under exponential-Euler discretization, this is equivalent to the real SSM

$$h_t = e^{\Delta_t a_t} \mathcal{R}_t h_{t-1} + \Delta_t \begin{bmatrix} B_t \\ \widehat{B}_t \end{bmatrix} x_t, \qquad y_t = \begin{bmatrix} C_t \\ -\widehat{C}_t \end{bmatrix}^\top h_t,$$

where $h_t = (\Re z_t, \Im z_t) \in \mathbb{R}^N$ and

$$\mathcal{R}_t = \mathrm{blockdiag}\{R(\Delta_t \theta_{t,j})\}_{j=1}^m, \qquad R(\vartheta) = \begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix}.$$

Proof. Appendix A.2.21.

Proposition 4.29 (Data-dependent RoPE form with exponential trapezoids; source propositions of Lahoti et al. [26])**.** *Let $\mathcal{R}_t$ be the block rotation from Proposition 4.28, and suppose the realified trapezoidal recurrence is*

$$h_t = \alpha_t \mathcal{R}_t h_{t-1} + \beta_t \mathcal{R}_t B_{t-1} x_{t-1} + \gamma_t B_t x_t.$$

Define $Q_t = \prod_{s=0}^{t} \mathcal{R}_s^\top$. Since the rotation blocks are orthogonal and commute across time,

$$\widetilde{h}_t := Q_t h_t$$

satisfies the scalar-transition recurrence

$$\widetilde{h}_t = \alpha_t \widetilde{h}_{t-1} + \beta_t Q_{t-1} B_{t-1} x_{t-1} + \gamma_t Q_t B_t x_t, \qquad y_t = (Q_t C_t)^\top \widetilde{h}_t.$$

Thus the complex transition can be moved from the recurrent state transition into cumulative, data-dependent rotations of the B and C projections.

Proof. Appendix A.2.22.

Proposition 4.30 (MIMO as SISO components, reformulating Lahoti et al. [26])**.** *Let ρ be the MIMO rank. Absorb the scalar input coefficient (Δ_t in the source two-term presentation, or γ_t in the notation above) into B_t. For $B_t = [B_t^{(1)}, \dots, B_t^{(\rho)}]$, $C_t = [C_t^{(1)}, \dots, C_t^{(\rho)}]$, and scalar input components $x_t^{(j)}$, define*

$$h_t^{(j)} = \alpha_t h_{t-1}^{(j)} + \gamma_t B_t^{(j)} x_t^{(j)}, \qquad h_t = \sum_{j=1}^{\rho} h_t^{(j)}, \qquad y_t^{(i)} = (C_t^{(i)})^\top h_t.$$

Then

$$y_t^{(i)} = \sum_{j=1}^{\rho} (C_t^{(i)})^\top h_t^{(j)} = \sum_{j=1}^{\rho} \mathrm{SSM}(\alpha, \gamma, B^{(j)}, C^{(i)}, x^{(j)})_t.$$

Equivalently, the matrix recurrence

$$H_t = \alpha_t H_{t-1} + B_t X_t^\top, \qquad Y_t = H_t^\top C_t$$

is an algebraic packing of ρ^2 SISO input-output components with shared transition sequence.

Proof. Appendix A.2.23.

Observation. *The source-derived Mamba-3 results are representation and discretization facts. In the proof map of this note, exponential trapezoids refine the selective recurrence, complex rotations connect to the eigenvalue and state-tracking results of Section 6, and the bidiagonal mask places Mamba-3 inside the SSD/semiseparable landscape of Section 5.*

4.6 Derived Compatibility Consequences

The following statements are consequences of the displayed Mamba-3 recurrences. They are not additional claims attributed to Lahoti et al. [26].

Proposition 4.31 (Derived affine scan compatibility)**.** *The exponential-trapezoidal recurrence remains an affine recurrent map. In the real scalar-transition case,*

$$h_t = \alpha_t h_{t-1} + b_t, \qquad b_t = \beta_t B_{t-1} x_{t-1} + \gamma_t B_t x_t.$$

In the realified complex case,

$$h_t = T_t h_{t-1} + b_t, \qquad T_t = \alpha_t \mathcal{R}_t, \qquad b_t = \beta_t \mathcal{R}_t B_{t-1} x_{t-1} + \gamma_t B_t x_t.$$

Consequently Proposition 4.3 applies after the state-input term b_t has been materialized.

Proof. Appendix A.2.24.

Proposition 4.32 (Derived off-diagonal semiseparable form)**.** *In the scalar-transition setting of Proposition 4.27, let $K_{t,s}$ be the coefficient of $v_s = B_s x_s$ in h_t. Then*

$$K_{s,s} = \gamma_s, \qquad K_{t,s} = 0 \quad (s > t),$$

and for $s < t$,

$$K_{t,s} = \left(\prod_{r=s+2}^{t} \alpha_r \right) \left(\alpha_{s+1}\gamma_s + \beta_{s+1} \right).$$

Thus the strict lower-triangular part is generated by one scalar SSM semiseparable recurrence. The extra Mamba-3 trapezoidal term changes the input injection weights, not the scalar propagation law below the diagonal.

Proof. Appendix A.2.25.

Proposition 4.33 (Derived rotation spectrum and parity sign)**.** *For one complex mode, the realified transition*

$$T_t = e^{\Delta_t a_t} R(\Delta_t \theta_t)$$

has eigenvalues

$$e^{\Delta_t a_t \pm i \Delta_t \theta_t}.$$

Hence positive-real-eigenvalue hypotheses do not apply unless $\Delta_t \theta_t \in 2\pi\mathbb{Z}$.

Moreover, for binary inputs $x_t \in \{0, 1\}$, choose $a_t = 0$, $\Delta_t = 1$, $\theta_t = \pi x_t$, no input injection, and $h_0 = (1, 0)^\top$. Then

$$h_T = ((-1)^{\sum_{t=1}^{T} x_t}, 0)^\top.$$

The linear readout $e_1^\top h_T$ represents the parity sign, and a final affine threshold recovers the parity bit.

Proof. Appendix A.2.26.

Proposition 4.34 (Derived finite abelian state tracking by Mamba-3 rotations)**.** *Let G be a finite abelian group, let Σ be a finite alphabet, and let $\varphi : \Sigma \to G$ be any symbol-to-state increment map. For a word $x_{1:T} \in \Sigma^T$, write*

$$g_T = \sum_{t=1}^{T} \varphi(x_t) \in G.$$

For every subset $A \subseteq G$, there is a zero-input-injection, realified complex Mamba-3 recurrence with $2|G|$ real state coordinates and input-dependent phases such that its final linear readout equals

$$\mathbf{1}_A(g_T).$$

Consequently a fixed threshold recognizes exactly the language

$$L_A = \{x_{1:T} : \sum_{t=1}^{T} \varphi(x_t) \in A\}.$$

Proof. Appendix A.2.27.

Proposition 4.35 (Derived componentwise lifting to MIMO)**.** *Let $\mathcal{T}_{ij}$ be the SISO sequence operator from input component $x^{(j)}$ to the contribution of output component $y^{(i)}$ in Proposition 4.30. Then the MIMO operator is*

$$y^{(i)} = \sum_{j=1}^{\rho} \mathcal{T}_{ij} x^{(j)}.$$

Consequently every exact linear identity for each SISO component lifts to the MIMO operator by summing over j. More quantitatively, if

$$\|\mathcal{T}_{ij} u\|_{\mathcal{V}} \le L_{ij} \|u\|_{\mathcal{U}}$$

for scalar-sequence norms $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{V}}$, then

$$\left(\sum_i \|y^{(i)}\|_{\mathcal{V}}^2 \right)^{1/2} \le \|L\|_{2\to 2} \left(\sum_j \|x^{(j)}\|_{\mathcal{U}}^2 \right)^{1/2}.$$

Proof. Appendix A.2.28.

Earlier result family	Transfer status	Mathematical reason
Selective scan	Applies	The update remains affine, $h_t = T_t h_{t-1} + b_t$, after materializing the trapezoidal state-input b_t.
SSD and semiseparable duality	Partly applies	The scalar-transition mask has a one-semiseparable strict lower part composed with a bidiagonal input mask; block rotations and MIMO require block/rank bookkeeping.
Positive-real spectrum limits	Does not apply verbatim	Complex Mamba-3 modes have eigenvalues $e^{\Delta_t a_t \pm i\Delta_t \theta_t}$, so the positive-real spectral hypothesis is violated except at zero phase.
Group-state constructions	Applies for finite abelian groups	Characters of finite abelian groups are realized by input-dependent complex phases; linear Fourier readout recovers any subset of the group state.
Diagonal formal-language limits	Needs restatement	Results for simultaneously diagonal real or positive transitions do not automatically cover real block rotations or complex modes.
SISO stability and norm bounds	Lifts with factors	Componentwise SISO bounds lift through MIMO with the matrix norm of the rank-coupling constants.
Copying and fixed-state lower bounds	Assumption dependent	The fixed-state bottleneck remains relevant, but the exact lower bound must match the Mamba-3 precision, state, and transition assumptions.

4.7 Derived Mamba-3 Transfer Theorems

Let $\mathcal{M}_{\mathrm{EE}}^{\mathrm{sc}}$ denote the class of scalar-decay exponential-Euler selective SSM sequence maps

$$h_t = e^{\Delta_t A_t} h_{t-1} + \Delta_t B_t x_t, \qquad y_t = C_t^\top h_t,$$

where $A_t \in \mathbb{R}$ is allowed to depend on the input and time. Let $\mathcal{M}_3$ denote the class generated by the Mamba-3 recurrences displayed above, including the scalar real, realified complex, and MIMO constructions. The following transfer statements are derived set-inclusion facts in this note; they are not source theorem claims of Lahoti et al. [26].

Proposition 4.36 (Derived exponential-Euler subfamily containment)**.** *The scalar-decay exponential-Euler class is contained in the Mamba-3 class:*

$$\mathcal{M}_{\mathrm{EE}}^{\mathrm{sc}} \subseteq \mathcal{M}_3.$$

More explicitly, every recurrence in $\mathcal{M}_{\mathrm{EE}}^{\mathrm{sc}}$ is obtained from Proposition 4.26 by choosing

$$\lambda_t = 1, \qquad \beta_t = 0, \qquad \gamma_t = \Delta_t,$$

and by using the zero-phase real subfamily of the complex representation.

Proof. Appendix A.2.29.

Proposition 4.37 (Derived existential theorem transfer)**.** *Let $\mathcal{P}$ be a property of sequence maps, and suppose an earlier positive result proves*

$$\exists f \in \mathcal{M}_{\mathrm{EE}}^{\mathrm{sc}} \quad \textit{such that} \quad \mathcal{P}(f).$$

Then the same existential statement holds for Mamba-3:

$$\exists f \in \mathcal{M}_3 \quad \textit{such that} \quad \mathcal{P}(f).$$

Thus constructive upper bounds, exact simulations, and expressivity constructions from the scalar-decay exponential-Euler subfamily transfer to Mamba-3 without changing the constructed sequence map.

Proof. Appendix A.2.30.

Proposition 4.38 (Derived universal theorem transfer criterion)**.** *Let $\mathcal{C}$ be a model class and let $\mathcal{Q}$ be a universal conclusion, for example an impossibility statement or a stability bound. If a theorem proves*

$$\forall f \in \mathcal{C}, \qquad \mathcal{Q}(f),$$

then it applies to all of Mamba-3 only under the additional inclusion $\mathcal{M}_3 \subseteq \mathcal{C}$. If instead only a restricted Mamba-3 subclass $\mathcal{M}_3' \subseteq \mathcal{C}$ is known, the theorem transfers only to $\mathcal{M}_3'$.

In particular, lower bounds whose hypotheses require positive real transition eigenvalues apply to the positive-real zero-phase restriction

$$\Delta_t \theta_t \in 2\pi\mathbb{Z}$$

of the realified Mamba-3 complex modes, but Proposition 4.33 shows that they do not cover the full rotation family.

Proof. Appendix A.2.31.

Remark. *Proposition 4.37 is the formal mechanism for importing positive constructions into Mamba-3. Proposition 4.38 is the corresponding check for limitations: the hypotheses, not the architecture name, determine whether a theorem applies.*

5 State-Space Duality and Attention

The common object in the attention-duality papers is a causal lower triangular sequence matrix. Recurrences, semiseparable matrices, and structured masked attention are different representations of this object under different structural assumptions.

5.1 Hidden Attention in Mamba-1

Proposition 5.1 (Hidden attention matrix; Ali et al. [1])**.** *Consider one S6 channel with $h_0 = 0$,*

$$h_i = \bar{A}_i h_{i-1} + \bar{B}_i x_i, \qquad y_i = C_i h_i.$$

Then $y = \widetilde{\alpha} x$, where $\widetilde{\alpha}$ is causal lower triangular with entries

$$\widetilde{\alpha}_{ij} = C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j, \qquad j \le i.$$

If $\bar{A}_i$ is diagonal with state dimension N, then each channel-level coefficient is the sum of N coordinatewise attention coefficients,

$$\widetilde{\alpha}_{ij} = \sum_{m=1}^{N} C_i[m] \bar{B}_j[m] \prod_{k=j+1}^{i} \bar{A}_k[m,m].$$

Proof. Appendix A.3.1.

Remark. *This is an exact unrolling identity for the S6 recurrence. The later query-history-key expression in Ali et al. [1] substitutes the Mamba parameter maps into this identity and uses an approximation to simplify Softplus terms.*

5.2 SSMs as Semiseparable Matrices

Definition 5.2 (N-semiseparable and N-SSS matrices; Dao and Gu [7])**.** A lower triangular matrix $M \in \mathbb{R}^{T \times T}$ is N-semiseparable if every submatrix contained on or below the diagonal has rank at most N. It has an N-sequentially semiseparable representation if there are $A_t \in \mathbb{R}^{N \times N}$ and $B_t, C_t \in \mathbb{R}^N$ such that

$$M_{ji} = C_j^\top A_j A_{j-1} \cdots A_{i+1} B_i, \qquad i \le j.$$

Theorem 5.3 (SSM equals SSS; Dao and Gu [7], Theorem 3.5)**.** *For state size N, set $M = \mathrm{SSS}(A, B, C)$. Then*

$$\mathrm{SSM}(A, B, C)(x) = Mx.$$

The matrix M is an N-SS matrix in sequentially semiseparable representation.

Proof. Appendix A.3.2.

Proposition 5.4 (Semiseparable matrix-vector multiplication; Dao and Gu [7], Proposition 3.6)**.** *An N-semiseparable matrix of size T has an $O(NT)$-parameter compressed representation and supports matrix-vector multiplication in $O(NT)$ time and space.*

Proof. Appendix A.3.3.

Theorem 5.5 (Efficient SSM computation; Dao and Gu [7], Theorem 3.7)**.** *Any SSM of state size N on sequence length T can be computed in $O(TN)$ time, not counting potential preprocessing.*

Proof. Appendix A.3.4.

5.3 Structured State-Space Duality

Proposition 5.6 (Scalar-identity SSD duality; Dao and Gu [7], Section 5)**.** *Suppose the SSM transition has scalar-identity form $A_t = a_t I_N$. Let*

$$L_{ji} = \prod_{k=i+1}^{j} a_k, \qquad i \le j,$$

and $L_{ji} = 0$ for $i > j$. If $B, C \in \mathbb{R}^{T\times N}$ collect the input and output projections, then the SSM sequence matrix is

$$M = L \odot (CB^\top).$$

Thus the same transformation has the recurrent SSM form and the quadratic structured masked-attention form.

Proof. Appendix A.3.5.

Corollary 5.7 (1-SS SMA as a diagonal SSM; Dao and Gu [7], Corollary 5.1)**.** *Structured masked attention with a 1-semiseparable mask is a special case of a diagonal SSM whose diagonal transition matrix is a scalar multiple of the identity.*

Proof. Appendix A.3.5.

Theorem 5.8 (Autoregressive attention is semiseparable; Dao and Gu [7], Theorem 5.2)**.** *For any structured masked-attention instance that is an autoregressive process of bounded order, the structured mask L must be semiseparable.*

Proof. Appendix A.3.6.

Remark. *The source duality is exact for scalar-identity SSMs and structured masked attention with a 1-semiseparable mask. More general semiseparable masked attention is a broader algebraic class; the standard scalar-identity SSM form does not cover it without changing the expansion/contraction representation.*

Theorem 5.9 (SSD algorithm; Dao and Gu [7], Theorem 6.1)**.** *For an SSD model with state expansion N and head dimension $P = N$, there is an algorithm that computes the model on any $X \in \mathbb{R}^{T\times P}$ using $O(TN^2)$ training FLOPs, $O(TN)$ inference FLOPs, $O(N^2)$ inference memory, and work dominated by matrix multiplications. In the block decomposition, the source chooses block length $Q = N$; the displayed bounds are unchanged by a final shorter block.*

Proof. Appendix A.3.7.

5.4 Diagonal SSD Beyond Scalar Identity

Proposition 5.10 (Diagonal SSM as a sum of 1-SS attention heads; Hu et al. [20], Section 4.1)**.** *Let M be the sequence matrix of a diagonal SSM,*

$$M_{ji} = c_j^\top A_j A_{j-1} \cdots A_{i+1} b_i, \qquad i \le j,$$

where each A_t is diagonal in $\mathbb{R}^{N\times N}$. Writing $a_t^{(n)} = (A_t)_{nn}$, $b_t^{(n)} = (b_t)_n$, and $c_t^{(n)} = (c_t)_n$, one has

$$M = \sum_{n=1}^{N} 1\,\mathrm{SS}(a^{(n)}) \odot \left(c^{(n)} (b^{(n)})^\top\right).$$

If every A_t is invertible, this representation can be reparameterized as a single 1-SS masked-attention dual

$$M = 1\,\mathrm{SS}(\mathbf{1}) \odot (C'B'^{\top}).$$

Proof. Appendix A.3.9.

Definition 5.11 (Fine 1-SS masks and new columns; Hu et al. [20])**.** Let $L = 1\,\mathrm{SS}(a_1, \ldots, a_T)$ be a 1-semiseparable mask. It is fine if $a_1 a_2 \cdots a_T \neq 0$. For a lower triangular matrix M, column t is new if the tail $M_{t:T,t}$ is not in the span of the previous tails $M_{t:T,s}$ with $s < t$. If a 1-SS mask is not fine, its zero transition entries split the lower triangular matrix into diagonal blocks.

Theorem 5.12 (General state-space duality condition; Hu et al. [20], Theorem 4.1)**.** *Let M be an N-semiseparable matrix corresponding to an SSM. This SSM has a 1-SS masked-attention dual if and only if M has several diagonal blocks containing all nonzero entries of M, and each diagonal block has at most N new columns.*

Proof. Appendix A.3.10.

5.5 Boundary of the Duality

Proposition 5.13 (Softmax rank obstruction; Hu et al. [20], Section 5)**.** *There are rank-one score matrices V for which* Softmax(V) *has full rank T, and every submatrix of* Softmax(V) *is full rank. Since a finite-state SSM dual would require a fixed semiseparable rank, standard softmax attention does not admit the same finite-dimensional SSD duality in general.*

Proof. Appendix A.3.10.

Proposition 5.14 (Low-state SSM obstruction; Hu et al. [20], Section 5)**.** *For every state dimension $N \geq 2$ and all sufficiently large T, there are SSM parameters whose sequence matrix is 2-semiseparable, for example $M = I_T + E^{T,1}$, but which have no representation*

$$M = L \odot (QK^{\top}), \qquad Q, K \in \mathbb{R}^{T \times N},$$

with a 1-SS mask L.

Proof. Appendix A.3.10.

Proposition 5.15 (Mamba S6 head pattern; Dao and Gu [7], Proposition 7.2)**.** *In the SSD attention-dual vocabulary, the selective SSM layer of Mamba has head dimension $P = 1$: each channel has independent SSM dynamics A, while B and C, corresponding to keys and queries in the attention dual, are shared across input channels.*

Proof. Appendix A.3.8.

Thus Mamba-2 is not only an implementation change. Its core SSD layer moves from Mamba-1's diagonal per-channel transition to scalar-identity transitions organized with larger head dimension, exposing the exact structured attention-dual form and the block SSD algorithm.

6 Expressivity and Formal Languages

Formal-language results give a controlled way to ask what an SSM layer can represent at arbitrary sequence lengths. The recurring objects are exact recognition, next-symbol prediction, and algebraic state tracking.

6.1 Recognition, Prediction, and State Tracking

Definition 6.1 (Recognition and predictive modeling, local convention)**.** Let $\mathcal{L} \subseteq \Sigma^*$ be a language. A sequence model recognizes $\mathcal{L}$ if its final output separates $w \in \mathcal{L}$ from $w \notin \mathcal{L}$. It predictively models $\mathcal{L}$ if, for every valid prefix w, it outputs the set of symbols

$$\{\sigma \in \Sigma : \exists v \in \Sigma^* \text{ such that } w\sigma v \in \mathcal{L}\}.$$

Definition 6.2 (Group state tracking, local convention)**.** Let G be a finite group or semigroup. Given inputs $x_1, \ldots, x_T \in G$, the target at time t is

$$g_t = x_1 x_2 \cdots x_t.$$

Parity, modular counting, cyclic automata, and permutation tracking are instances of this object.

6.2 Star-Free Languages and Counters

In the results of Sarrof et al. [35], finite precision means an unbounded number of integer bits but only a fixed number of fractional bits, independent of input length.

Theorem 6.3 (Flip-Flop; Sarrof et al. [35], Theorem 1)**.** *There is a two-layer finite-precision SSM that predictively models the Flip-Flop language L_{FF} at all input lengths.*

Proof. Appendix A.4.1.

Theorem 6.4 (Star-free characterization; Sarrof et al. [35], Theorem 4)**.** *Let $\mathcal{L}$ be a regular language. For the finite-precision SSM abstraction of Sarrof et al. [35] satisfying the nonnegative-gate assumption, the following are equivalent:*

(i) *$\mathcal{L}$ is predictively modeled at all input lengths,*

(ii) *$\mathcal{L}$ is star-free.*

Proof. Appendix A.4.2.

Theorem 6.5 (Counter languages and bounded Dyck; Sarrof et al. [35], Theorems 5 and 6)**.** *Under the finite-precision convention of Sarrof et al. [35], with bounded fractional precision and unbounded integer bits, the languages*

$$\mathrm{Dyck}_1, \quad \text{Shuffle-Dyck}_k, \quad \textit{n-ary Boolean expressions}, \quad a^n b^n, \quad a^n b^n c^n, \quad a^n b^n c^n d^n$$

can each be predictively modeled by an SSM. Moreover, for bounded-depth Dyck, which uses only bounded counter values, there is a two-layer SSM with width $O(h \log K)$ that predictively models $\mathrm{Dyck}_{K,h}$ at all input lengths.

Proof. Appendix A.4.3.

6.3 Finite-Context Prediction

Theorem 6.6 (n-gram prediction; Nandakumar et al. [30], Theorem 4.1)**.** *Let $\mathcal{L}$ be a language over a vocabulary W, and let f^*_{ng} be an n-gram language model obtained from a finite rule datum (P, f_{ng}). For every $\epsilon > 0$, there is a state-space language model f^*_S such that f^*_S and f^*_{ng} are ϵ-equivalent on $\mathcal{L}$. The SSM component may be chosen with one hidden layer of size $|P|$, and the embedding dimension may be any fixed $e \geq 1$.*

Proof. Appendix A.4.4.

Proposition 6.7 (Memorization and finite context; Nandakumar et al. [30], Propositions 4.4 and 4.5)**.** *Let $\{(x_i, y_i)\}_{i=1}^K$ be distinct finite input-output sequence pairs, where each x_i has length n. There is a one-hidden-layer SSM with K hidden units that memorizes these pairs, and its transition matrix may be chosen so that $A^n = 0$. Conversely, if an SSM transition satisfies $A^n = 0$, then the output at time T depends only on $x_{T-n+1}, \ldots, x_T$.*

Proof. Appendix A.4.4.

6.4 Polynomial Selectivity

Proposition 6.8 (S6 versus linear attention; Cohen-Karlik et al. [5], Lemmas 1 and 2)**.** *For input length $L \geq 3$, there is a function $f : \mathbb{R}^L \to \mathbb{R}$ implemented by one channel of the simplified S6 layer analyzed by Cohen-Karlik et al. [5] such that a single-head causal linear-attention model requires logarithmically many stacked layers to express f. Conversely, every function expressible by a single causal linear-attention head is expressible by one channel of the same S6 class.*

Proof. Appendix A.4.5.

Lemma 6.9 (Monomial extraction by three Mamba layers; Cohen-Karlik et al. [5], Lemma 3)**.** *For a scalar input sequence $x \in \mathbb{R}^L$, a three-layer simplified Mamba model with sufficiently many channels, learnable positional encodings, and a linear encoder in the first layer can isolate any position j and output*

$$c\, x_j^P$$

in a prescribed channel, for any coefficient $c \in \mathbb{R}$ and any monomial degree P in the polynomial under consideration. Sufficient zero padding is used when the construction needs more positions than the original sequence length.

Proof. Appendix A.4.5.

Theorem 6.10 (Polynomial universality of stacked Mamba blocks; Cohen-Karlik et al. [5], Theorem 2)**.** *For a scalar input sequence $x \in \mathbb{R}^L$, four stacked Mamba layers, with sufficiently many channels, learnable positional encodings, sufficient padding, and a linear encoder in the first layer, can express any multivariate polynomial of bounded degree in x. This is in the paper's simplified setting with state size $N = 1$ and the SiLU activations removed.*

Proof. Appendix A.4.5.

6.5 Finite Automata and Diagonal State Tracking

Proposition 6.11 (FSA-to-selective-SSM mapping; Terzić et al. [40])**.** *Let $\mathcal{A} = (Q, \Sigma, q_{\text{init}}, \delta)$ be a deterministic finite automaton. With one-hot state encodings e_q, define*

$$A(\sigma) = \sum_{q \in Q} e_{\delta(q,\sigma)} e_q^\top, \qquad b(\sigma) = 0.$$

Then the selective recurrence

$$\mathbf{q}_t = A(x_t)\mathbf{q}_{t-1}, \qquad \mathbf{q}_0 = e_{q_{\text{init}}},$$

exactly emulates the automaton state sequence.

Proof. Appendix A.4.6.

Definition 6.12 (PD-SSM transition class, Terzić et al. [41])**.** Let $\mathbb{H}^{N\times N}$ be the set of matrices with exactly one nonzero entry in each column. A PD-SSM transition has the form

$$A_t = P_t D_t, \qquad P_t \in \mathbb{H}^{N\times N}, \qquad D_t = \operatorname{diag}(d_{t,1}, \ldots, d_{t,N}) \in \mathbb{C}^{N\times N}.$$

Proposition 6.13 (One-hot-column monoid and linear multiplication, Properties 1–2 of Terzić et al. [41])**.** *The set $\mathbb{H}^{N\times N}$ is a monoid under matrix multiplication. Moreover, if $A, B \in \mathbb{H}^{N\times N}$, then $C = AB \in \mathbb{H}^{N\times N}$ can be computed in $\Theta(N)$ arithmetic operations.*

Proof. Appendix A.4.7.

Proposition 6.14 (PD-SSM stability, Terzić et al. [41])**.** *Let*

$$h_t = A_t h_{t-1} + b_t, \qquad A_t = P_t D_t, \qquad P_t \in \mathbb{H}^{N\times N},$$

and assume $\|A_t\|_\infty \le 1 - \varepsilon$ for some $\varepsilon \in (0, 1]$. If $\|h_0\|_2 \le B$ and $\|b_t\|_2 \le B$ for all t, then

$$\|h_t\|_2 \le \frac{\sqrt{N}\, B}{\varepsilon} \qquad \text{for all } t.$$

Proof. Appendix A.4.8.

Theorem 6.15 (PD-SSM finite-state tracking, Propositions 1–3 of Terzić et al. [41])**.** *Every deterministic FSA with N states can be exactly represented by a single-layer PD-SSM with state size N and a linear readout of size $N \times N$. Under unique state encodings, for every N there is an N-state FSA that cannot be emulated by any single-layer SSM with state size less than $N-1$. If arbitrary precision and a lookup readout are allowed, then a scalar recurrence of the form*

$$x_{t+1} = x_t + u_t \sqrt{p_t},$$

with p_t the tth prime and rational input encodings u_t, can encode any FSA into a scalar state.

Proof. Appendix A.4.9.

Proposition 6.16 (Commutativity of simultaneously diagonal transitions; Terzić et al. [40], Proposition 1)**.** *Under the FSA mapping above, suppose $b(x_t) = 0$ and all transition matrices $A(x_t)$ are simultaneously diagonalizable. Then the resulting single-layer selective SSM can emulate only commutative automata.*

Proof. Appendix A.4.6.

Theorem 6.17 (Diagonal complex group tracking; Shakerinava et al. [36], Theorems 1 and 2)**.** *Let G be a finite group. For the input-dependent complex-valued diagonal SSM abstraction at finite precision, with the source assumptions that A, b, and the decoder are universal function approximators:*

$$\textit{one layer tracks } G \quad \Longleftrightarrow \quad G \textit{ is Abelian.}$$

More generally, a k-layer DCD SSM tracks G if and only if there exists a subnormal series

$$G = G_k \trianglerighteq G_{k-1} \trianglerighteq \cdots \trianglerighteq G_1 \trianglerighteq G_0 = \{e\}$$

such that every quotient G_{i+1}/G_i is Abelian.

Proof. Appendix A.4.10.

6.6 Eigenvalues and Regular Languages

Theorem 6.18 (Eigenvalue obstructions; Grazzi et al. [9], Theorems 1 and 2)**.** *Let an LRNN have finitely many finite-precision layers of the form $H_t = A(x_t)H_{t-1} + B(x_t)$. If it solves parity at arbitrary input lengths, then in at least one layer there is an input x such that $A(x)$ has an eigenvalue outside $\mathbb{R}_{\geq 0}$. If an L-layer LRNN counts modulo m, equivalently recognizes $(1^m)^*$, where m is not a power of two, then there exist $i \in \{1, \ldots, L\}$ and inputs $x_1, \ldots, x_{2^{i-1}}$ such that*

$$A_i(x_1)A_i(x_2)\cdots A_i(x_{2^{i-1}})$$

has a nonreal eigenvalue.

Proof. Appendix A.4.11.

Theorem 6.19 (Generalized Householder positive construction; Grazzi et al. [9], Theorems 3 and 4)**.** *Let $M_k^n(\Omega)$ denote products of k generalized Householder matrices in $\mathbb{R}^{n\times n}$ whose nontrivial eigenvalues lie in Ω. Every finite automaton whose transition monoid is a group can be implemented by a one-layer finite-precision LRNN with transitions in $M_{k-1}^n(\{-1, 1\})$, where the transition monoid embeds in a subgroup of S_n and k is the maximum number of points displaced by a one-symbol transition. With repeated products of generalized Householder matrices whose eigenvalues lie in $[-1, 1]$, LRNNs can recognize every regular language.*

Proof. Appendix A.4.11.

Theorem 6.20 (Adaptive unitary expressivity; Karuvally et al. [24], Lemmas 1–2 and Theorem 2)**.** *Under the paper's logarithmically bounded precision assumption, a single-layer AUSSM can count modulo k for every $k \in \mathbb{Z}_+$, hence can simulate arbitrary cyclic group automata. Interleaved Mamba and AUSSM blocks can implement cascade products of automata simulated by the individual blocks. Consequently, hybrid Mamba+AUSSM can recognize every solvable regular language, i.e. every regular language whose syntactic monoid has no non-solvable subgroup.*

Proof. Appendix A.4.12.

6.7 Placement of Dense Selective Transitions

The Selective Dense SSM of Terzić et al. [40] replaces diagonal transition generation by a softmax-weighted dictionary of dense transition matrices, followed by operator normalization and a linear readout. The empirical part studies finite-state emulation and length generalization for this architecture. The simultaneously diagonal proposition above is the theorem-level limitation used here.

7 Limitations

The limitation results use different mathematical abstractions: fixed-state copying, uniform circuit simulations, finite-precision eigenvalue dynamics, and algebraic restrictions on diagonal group tracking.

7.1 Copying

Theorem 7.1 (Copying separation; Jelassi et al. [22], Theorems 2.3 and 2.7)**.** *For every n, there is a depth-2 Transformer T of dimension $O(n \log D)$ such that, for all $2n \leq L \leq D^n$ and every length-L copy distribution $\mathcal{D}_L$,*

$$\mathrm{err}_{\mathcal{D}_L}(T) < p_{n\text{-}gram}(\mathcal{D}_L),$$

where $p_{n\text{-}gram}$ is the probability of an n-gram collision in the input. In contrast, for every GSSM H with finite state space S and the uniform length-L copy distribution,

$$\mathrm{err}_{\mathcal{D}_L}(H) > 1 - |S|D^{-L}.$$

Hence if $\mathrm{mem}(S) < L \log D - 1$, *then* $\mathrm{err}_{\mathcal{D}_L}(H) > 1/2$.

Proof. Appendix A.5.1.

This is a memory-recovery result, not a formal-language state-tracking result. Copying a random string and tracking a finite algebraic state are different tasks.

7.2 Circuit-Complexity Containment

Theorem 7.2 (Log-precision SSMs in TC^0; Merrill et al. [29], Theorems 4.1, 4.3, and 4.5)**.** *For log-precision generalized linear SSMs of bounded depth, the convolutional form is computable by an* L*-uniform* TC^0 *circuit family in each of the following cases:*

(i) $\bar{A}_i, \bar{B}_i, C_i, D_i$ *are independent of* i,

(ii) $\bar{A}_i = \mathrm{diag}(\bar{a}_i)$ *and all tokenwise parameters are* TC^0*-computable,*

(iii) $\bar{A}_i = W\,\mathrm{diag}(\bar{a}_i)W^{-1}$ *for a fixed* W.

Consequently, S4 and the S6 layer used by Mamba satisfy the stated TC^0 *containment in this formal model [29, Corollaries 4.2 and 4.4].*

Proof. Appendix A.5.2.

Corollary 7.3 (Merrill et al. [29], Corollary 4.6)**.** *Assuming* $\mathrm{TC}^0 \neq \mathrm{NC}^1$*, no log-precision bounded-depth SSM with the S4 or S6 architecture solves the word problem for* S_5*, or any other* NC^1*-hard problem, under the paper's formalization.*

Proof. Appendix A.5.2.

Theorem 7.4 (DLOGTIME-uniform simulation of Selective SSMs and Mamba; Chen et al. [3], Theorems 4.4, 4.5, and C.22)**.** *Let the sequence length, width, hidden dimension, and floating-point precision be bounded by* $\mathrm{poly}(n)$*. The Selective SSM map and the Mamba block map formalized by Chen et al. [3] are computable by DLOGTIME-uniform threshold circuits of constant depth and polynomial size. If* $\mathrm{TC}^0 \neq \mathrm{NC}^1$*, then constant-depth Selective SSMs and Mamba blocks with hidden dimension* $O(n)$ *cannot solve arithmetic formula evaluation, Boolean formula value, or permutation-composition problems in that model.*

Proof. Appendix A.5.3.

7.3 Metric Automata and Robust Finite Abstractions

Definition 7.5 (Metric automata objects, Dankowiakowski and Ronca [6])**.** A set $X \subseteq \mathbb{R}^d$ or $X \subseteq \mathbb{C}^d$ is η-finite if it is a finite union of compact path-connected components. A dynamics $D = \langle X, U, f \rangle$ is η-finite if both X and U are η-finite. Its canonical semiautomaton quotients X and U by membership in η-components. For $\varepsilon > 0$, D is ε-robust in an ambient space Ω if

$$B_\Omega(f(x,u),\varepsilon) \subseteq X \qquad \text{for every } x \in X,\ u \in U.$$

It is strongly ε-robust if, additionally, each state η-component contains an Ω-ball of radius at least ε.

Theorem 7.6 (Canonical automaton and finite datatype implementation, Theorems 1–2 of Dankowiakowski and Ronca [6])**.** *An η-finite system can implement exactly the same functions as its canonical finite automaton; these functions are necessarily regular. If the system has strongly ε-robust dynamics, then it can be implemented with floating-point operations at sufficiently high finite precision.*

Proof. Appendix A.5.4.

Theorem 7.7 (Robust LRD and Mamba FLIP-FLOP obstruction, Theorems 3–4 and 7 of Dankowiakowski and Ronca [6])**.** *Let an η-finite linear recurrent dynamics have at least two canonical states and an input inducing the identity transformation on the canonical semiautomaton. Then it is not ε-robust for any $\varepsilon > 0$. Consequently, a cascade of η-finite ε-robust LRDs cannot implement FLIP-FLOP. Under the source's Mamba parameterization and finite/robust hypotheses, SSMs with Mamba parameterization cannot recognize FLIP-FLOP.*

Proof. Appendix A.5.5.

Theorem 7.8 (Aperiodicity and star-free boundary, Theorem 6 of Dankowiakowski and Ronca [6])**.** *Let D be an η-finite LRD whose state-transition gates have only nonnegative eigenvalues. Then D is aperiodic in the metric-automata sense. Equivalently, its canonical semiautomaton is group-free. Cascades of such dynamics, including the source's finite-precision Mamba-style cascades under the stated nonnegative-eigenvalue hypotheses, can recognize only star-free regular languages.*

Proof. Appendix A.5.6.

Theorem 7.9 (Geometric Mamba results, Theorems 8–9 of Dankowiakowski and Ronca [6])**.** *As geometrically constrained systems, SSMs with Mamba parameterization can recognize all star-free languages. Conversely, if an η-finite LRD has nonnegative-eigenvalue transition gates and the covering is convex-regular, then the induced dynamics are aperiodic with respect to the covering; hence alternating functions cannot be implemented in that GCS setting.*

Proof. Appendix A.5.7.

7.4 Parity, Modularity, and Diagonal Dynamics

Theorem 7.10 (Nonnegative parity obstruction; Sarrof et al. [35], Theorem 2)**.** *No finite-precision SSM satisfying the nonnegative-gate assumption of Sarrof et al. [35], with the layerwise linear, GLU, or SwiGLU operations covered by their proof, can recognize PARITY at arbitrary input lengths. In particular, the authors identify Mamba-style positive diagonal gating as an instance of this obstruction under their surveyed assumptions.*

Proof. Appendix A.4.2.

Theorem 7.11 (Input dependence and negative eigenvalues must coincide; Khavari et al. [25], Theorem 4.4)**.** *A finite-precision LRNN consisting of diagonal S4D layers and positive diagonal Mamba layers, with skip connections and learnable initial states allowed, cannot solve parity. More generally, in the source's diagonal finite-precision setting, the proof applies to combinations of time-invariant layers whose complex phases have a common finite period and input-dependent nonnegative layers.*

Proof. Appendix A.5.8.

The theorem complements the eigenvalue obstructions of Grazzi et al. [9]: it is not enough for input dependence and negative or complex eigenvalues to appear somewhere in a diagonal stack; for parity, the analyzed setting needs them in the same recurrence layer.

7.5 Diagonal Selectivity and Algebraic Order

The diagonal state-tracking theorem of Shakerinava et al. [36], stated in the previous section, gives a sharper algebraic boundary inside the DCD abstraction: one layer tracks exactly Abelian groups, and k layers track exactly groups with a length-k subnormal series of Abelian quotients. This is distinct from the circuit-containment results above: it is an exact group-theoretic classification for diagonal complex input-dependent recurrences at finite precision.

The simultaneously diagonal proposition of Terzić et al. [40] is narrower but useful. With $b = 0$ and a common diagonalizing basis, transition products commute, so order-sensitive automata cannot be represented by that particular single-layer mapping.

7.6 Relations

The copying theorem concerns recovery of long random strings from fixed state. The circuit theorems concern uniform low-depth simulation under explicit precision and depth assumptions. The eigenvalue theorems concern oscillatory dynamics needed for parity and modular counting. The DCD theorem concerns exact all-length group tracking by diagonal input-dependent dynamics. These are adjacent but not interchangeable statements.

8 Stability and Generalization

The results in this section are not one theorem family. They concern stable approximation, data-dependent uniform convergence, PAC bounds for deep stable SSMs, covering-number bounds for selective SSMs, and gradient descent on a simplified Mamba block. The shared variable is stability: old inputs, perturbations, covers, and gradients all accumulate through transition products.

Common notation. This section keeps source constants when they carry theorem scope. The letter T denotes a sequence horizon or continuous-time interval unless a result explicitly says otherwise. Sample sizes are denoted by n, m, or M according to the source theorem. The spectral abscissa of a continuous matrix A is

$$s(A) = \max_{\lambda \in \operatorname{spec}(A)} \Re\lambda.$$

Stability means $s(A) < 0$ in continuous time, or a horizon-uniform system norm bound in discrete time. In the final training-dynamics subsection, L is the sequence length and T_{gd} is the number of gradient descent iterations.

8.1 Stable Approximation and Memory Decay

Definition 8.1 (Stable approximation objects; Wang and Li [42])**.** For a bounded Heaviside input $u^x(t) = x\mathbf{1}_{\{t \geq 0\}}$, the memory function of a bounded, causal, continuous, regular, time-homogeneous functional sequence $\mathbf{H} = \{H_t : t \in \mathbb{R}\}$ is

$$\mathcal{M}(\mathbf{H})(t) = \sup_{x \neq 0} \frac{\left|\frac{d}{dt} H_t(u^x)\right|}{\|x\|_\infty + 1}.$$

The approximation norm used below is

$$\|\mathbf{H} - \widehat{\mathbf{H}}\|_{W^{1,\infty}} = \sup_t \left(\|H_t - \widehat{H}_t\|_\infty + \left\| \frac{dH_t}{dt} - \frac{d\widehat{H}_t}{dt} \right\|_\infty \right).$$

For approximants $\widehat{\mathbf{H}}(\cdot, \theta_m)$ with $\theta_m = (\Lambda, U, b, c)$, let

$$E_m(\beta) = \sup_{\|\widetilde{\theta}_m - \theta_m\|_2 \leq \beta} \|\mathbf{H} - \widehat{\mathbf{H}}(\cdot, \widetilde{\theta}_m)\|_{W^{1,\infty}}, \qquad E(\beta) = \limsup_{m \to \infty} E_m(\beta).$$

The sequence $\{\widehat{\mathbf{H}}(\cdot, \theta_m)\}$ is a β_0-stable approximation of $\mathbf{H}$ if $E(0) = 0$ and E is continuous on $[0, \beta_0]$.

Theorem 8.2 (Curse of memory for unreparameterized SSMs; Wang and Li [42], Theorem 3.3)**.** *Let $\mathbf{H}$ be a sequence of bounded, causal, continuous, regular, and time-homogeneous functionals on $\mathcal{X}$ with decaying memory. Suppose that $\{\widehat{\mathbf{H}}(\cdot, \theta_m)\}_{m \geq 1}$ is a sequence of state-space models that β_0-stably approximates $\mathbf{H}$ in the preceding $W^{1,\infty}$ norm. Assume the hidden states are uniformly bounded, the activation is strictly increasing, continuously differentiable, and L_0-Lipschitz, and the model weights are uniformly bounded with $\theta_{\max} = \sup_m \|\theta_m\|_2 < \infty$. Then for every $0 < \beta < \beta_0$,*

$$\mathcal{M}(\mathbf{H})(t) \leq (d+1) L_0 \theta_{\max}^2 e^{-\beta t}, \qquad t \geq 0.$$

For an ℓ-layer SSM, the induced bound has the form

$$\mathcal{M}(\mathbf{H})(t) \leq (d+1) L_0^\ell \theta_{\max}^{\ell+1} P(t) e^{-\beta t}, \qquad t \geq 0,$$

where P is a polynomial of degree at most $\ell - 1$.

Proof. Appendix A.6.2.

Definition 8.3 (Stable reparameterization; Wang and Li [42], Definition 3.4)**.** A scalar continuous-time reparameterization $f : \mathbb{R} \to \mathbb{R}$ is stable if there is a continuous $g : [0, \infty) \to [0, \infty)$ with $g(0) = 0$ such that

$$\sup_w |f(w)| \sup_{|\widetilde{w} - w| \leq \beta} \int_0^\infty \left| e^{f(\widetilde{w})t} - e^{f(w)t} \right| dt \leq g(\beta).$$

Examples include $f(w) = -e^w$ and $f(w) = -\log(1 + e^w)$.

Theorem 8.4 (Stable approximation under stable reparameterization; Wang and Li [42], Theorem 3.5)**.** *Let* $\mathbf{H}$ *be any bounded, causal, continuous, regular, time-homogeneous linear functional. If* $\mathbf{H}$ *is approximated by a sequence of linear RNNs* $\{\widehat{\mathbf{H}}(\cdot,\theta_m)\}_{m\geq 1}$ *with stable reparameterization, then under the recurrent-weight perturbation model*

$$\widetilde{\lambda}_i = f(\widetilde{w}_i), \qquad |\widetilde{w}_i - w_i| \leq \beta,$$

with the input and readout coefficients fixed during this stability test, the approximation is a stable approximation.

Proof. Appendix A.6.3.

Proposition 8.5 (Gradient scale induced by a stability map; Wang and Li [42], Theorem 3.6)**.** *Suppose a diagonal recurrent eigenvalue is parameterized as* $\lambda = f(w)$*, where* w *is the trainable scalar. For a fixed target sequence* $\mathbf{H}$ *and approximating SSM sequence* $\widehat{\mathbf{H}}_m$*, the gradient with respect to* w *obeys*

$$G_f(w) := \left|\frac{\partial \,\mathrm{Loss}}{\partial w}\right| \leq C_{\mathbf{H},\widehat{\mathbf{H}}_m} \frac{|f'(w)|}{f(w)^2}$$

in continuous time. In discrete time,

$$G_f^D(w) \leq C_{\mathbf{H},\widehat{\mathbf{H}}_m} \frac{|f'(w)|}{(1-f(w))^2}.$$

The balanced-gradient criterion $C_{\mathbf{H},\widehat{\mathbf{H}}_m}|f'(w)|/f(w)^2 = L|w|$ *leads to*

$$f(w) = -\frac{1}{aw^2+b}$$

in continuous time, and to $f(w) = 1 - \frac{1}{aw^2+b}$ *in discrete time.*

Proof. Appendix A.6.3.

8.2 Data-Dependent Generalization for LTI SSMs

Theorem 8.6 (Data-dependent uniform convergence; Liu and Li [28], Theorem 1)**.** *Consider the single-block continuous-time linear SSM functional, with no skip term,*

$$x \longmapsto \int_0^T \rho_\theta(T-s)x(s)\,ds.$$

Let $\mu(t)$ *and* $K(s,t)$ *be the mean and covariance of the input process, and assume the normalized process*

$$\widetilde{x}(t) = \frac{x(t)-\mu(t)}{\sqrt{K(t,t)}}$$

is almost surely Holder continuous and pointwise sub-Gaussian as in the source Assumption 1. For a parameter set Θ*, define*

$$\psi(\Theta) = \sup_{\theta\in\Theta}\int_0^T |\rho_\theta(T-s)|\sqrt{K(s,s)}\,ds + \sup_{\theta\in\Theta}\left|\int_0^T \rho_\theta(T-s)\mu(s)\,ds\right|.$$

Then for every $\delta \in (0,1)$*, with probability at least* $1-\delta$ *over* n *training sequences,*

$$\sup_{\theta\in\Theta}|R_x(\theta) - R_n(\theta)| \leq (\psi(\Theta)+1)^2 \cdot O\left(\frac{\log^{3/2}(Tn/\delta)}{\sqrt{n}}\right),$$

where the hidden constant depends only on the Holder and sub-Gaussian process constants in Assumption 1.

Proof. Appendix A.6.4.

Proposition 8.7 (Initialization by the generalization proxy; Liu and Li [28], Proposition 1)**.** *Let* $\theta = (C,A,B)$ *and* $\rho_\theta(s) = Ce^{As}B$*. Define*

$$\tau(\theta) = \left(\int_0^T |\rho_\theta(T-s)|\sqrt{K(s,s)}\,ds + \left|\int_0^T \rho_\theta(T-s)\mu(s)\,ds\right|\right)^2$$

and rescale

$$\widetilde{C} = \frac{C}{\sqrt{\tau(\theta)}}, \qquad \widetilde{\theta} = (\widetilde{C}, A, B).$$

If the input process satisfies the theorem's process assumption, then

$$\mathbb{E}_x \left[\left| \int_0^T \rho_{\widetilde{\theta}}(T-s) x(s)\, ds \right| \right] \le O(\sqrt{\log T}),$$

with a constant depending only on the Holder and sub-Gaussian process constants.

Proof. Appendix A.6.4.

8.3 Length-Independent Bounds for Deep Stable SSMs

Definition 8.8 (Rademacher contraction; Rácz et al. [33], Definition 5.1)**.** Let $\Phi = \{\varphi : X_1 \to X_2\}$ be a class of maps between subsets of Banach spaces $\mathcal{X}_1, \mathcal{X}_2$. It is (μ, c)-Rademacher contractive if, for every $M \ge 1$ and every $Z \subseteq X_1^M$,

$$\mathbb{E}_\sigma \left[\sup_{\varphi \in \Phi} \sup_{\{u_i\}_{i=1}^M \in Z} \left\| \frac{1}{M} \sum_{i=1}^M \sigma_i \varphi(u_i) \right\|_{\mathcal{X}_2} \right] \le \mu\, \mathbb{E}_\sigma \left[\sup_{\{u_i\}_{i=1}^M \in Z} \left\| \frac{1}{M} \sum_{i=1}^M \sigma_i u_i \right\|_{\mathcal{X}_1} \right] + \frac{c}{\sqrt{M}}.$$

Lemma 8.9 (Composition of Rademacher contractions; Rácz et al. [33], Lemma 5.2)**.** *If Φ_1 is (μ_1, c_1)-Rademacher contractive and Φ_2 is (μ_2, c_2)-Rademacher contractive, then*

$$\Phi_2 \circ \Phi_1 = \{\varphi_2 \circ \varphi_1 : \varphi_1 \in \Phi_1, \varphi_2 \in \Phi_2\}$$

is $(\mu_1\mu_2, \mu_2 c_1 + c_2)$-Rademacher contractive.

Proof. Appendix A.6.5.

Theorem 8.10 (Deep stable SSM PAC bound; Rácz et al. [33], Theorem 5.5)**.** *Let $\mathcal{F}$ be the family of deep SSM models satisfying the source Assumption 4.7: fixed depth ℓ, scalar output, Lipschitz loss, bounded input and labels, stable SSM blocks with bounded ℓ_1 and $\mathcal{H}_2$ norms, bounded encoder/decoder weights, and bounded MLP or GLU blocks with the source activation and norm restrictions. Then, for data sample $S \sim \mathcal{D}^M$, with probability greater than $1-\delta$, every $f \in \mathcal{F}$ satisfies*

$$\mathcal{L}(f) - \mathcal{L}^S_{\mathrm{emp}}(f) \le \frac{\mu K_u L_\ell + c L_\ell}{\sqrt{M}} + K_\ell \sqrt{\frac{2\log(4/\delta)}{M}}.$$

Here $K_\ell = 2L_\ell \max\{K_{\mathrm{Dec}} r_\ell, K_y\}$, the radius sequence is

$$r_i = \begin{cases} K_{\mathrm{Enc}} K_u, & i = 1, \\ \widehat{r}_1(K_2 r_1) + \alpha_1 r_1, & i = 2, \\ \widehat{r}_{i-1}(K_1 r_{i-1}) + \alpha_{i-1} r_{i-1}, & i > 2, \end{cases}$$

and μ, c are the products and sums of the component contraction constants from Rácz et al. [33]:

$$\mu = K_{\mathrm{Enc}} K_{\mathrm{Dec}} (\mu_1 K_2 + \alpha_1) \prod_{i=2}^{\ell} (\mu_i K_1 + \alpha_i),$$

$$c = K_{\mathrm{Dec}} \sum_{j=1}^{\ell} \left[\prod_{i=j+1}^{\ell} (\mu_i K_1 + \alpha_i) \right] c_j.$$

None of these terms depends on the sequence length T.

Proof. Appendix A.6.5.

8.4 Selective SSM Bounds Through Attention

Theorem 8.11 (Selective SSM covering bound; Honarpisheh et al. [19], Theorem 3.3)**.** *Let $S = \{(u_{(i)}, z_{(i)})\}_{i=1}^m$ and let $\mathcal{F}_{\mathrm{SSM}}$ be the selective SSM block class in the source. Assume bounded inputs, bounded parameter norms for W_B, W_C, w, q, bounded and Lipschitz loss, and $\|A_c\|_2 \le \mathfrak{B}_A$, $\|A_c\|_{2,1} \le \mathfrak{M}_A$. Then with probability more than $1-\delta$, every $h \in \mathcal{F}_{\mathrm{SSM}}$ satisfies*

$$\left|\mathbb{E}_{u,z}\ell(h(u),z) - \frac{1}{m}\sum_{i=1}^m \ell(h(u_{(i)}), z_{(i)})\right| \le \frac{12\mathfrak{l}_\ell \mathcal{C}_{\mathcal{F}_{\mathrm{SSM}}}}{\sqrt{m}}\left(1 + \log\frac{\mathfrak{c}_\ell\sqrt{m}}{3\mathcal{C}_{\mathcal{F}_{\mathrm{SSM}}}}\right) + 3\mathfrak{c}_\ell\sqrt{\frac{\log(2/\delta)}{2m}},$$

where

$$\mathcal{C}_{\mathcal{F}_{\mathrm{SSM}}} = \widetilde{O}\Big(\mathfrak{M}_\Delta \mathfrak{B}_w \mathfrak{B}_u^3 \mathfrak{B}_B \mathfrak{B}_C \mathfrak{B}_A S_2 \left(\mathfrak{M}_\Delta^{2/3} N^{1/3} d^{1/3} + \mathfrak{B}_q^{2/3}\mathfrak{B}_u^{2/3}\right)^{3/2}\Big),$$

$$S_2 = \sum_{t=0}^{T-1} t\rho_A^t = \frac{\rho_A(1-\rho_A^T)}{(1-\rho_A)^2} - \frac{T\rho_A^T}{1-\rho_A}, \qquad \rho_A = (1+e^{p-\mathfrak{B}_q\mathfrak{B}_u})^{s_A+\eta}, \qquad \mathfrak{M}_\Delta = \log(1+e^{p+\mathfrak{B}_q\mathfrak{B}_u}).$$

Here s_A is the spectral abscissa of A_c and $\eta > 0$ is chosen so $s_A+\eta \neq 0$; at $\rho_A = 1$ the expression for S_2 is understood by its limit $T(T-1)/2$. The displayed capacity uses the source's final simplifying convention $\mathfrak{M}_{(\cdot)} = \mathfrak{B}_{(\cdot)}$ for all parameter classes except Δ.

Proof. Appendix A.6.6.

Proposition 8.12 (Linear-attention specialization; Honarpisheh et al. [19], Proposition 3.4)**.** *Under the source constant-step specialization $q = 0$, $p = e$, which the paper normalizes as $\Delta[t] = 1$, and $A_c = 0$, so that $A[t] = I$, the selective SSM reduces to causal linear attention. For the resulting class $\mathcal{F}_{\mathrm{LA}}$ parameterized by $\{W_C, W_B, w\}$, the same generalization bound holds with*

$$\mathcal{C}_{\mathcal{F}_{\mathrm{LA}}} = \widetilde{O}\big(T\mathfrak{B}_w\mathfrak{B}_B\mathfrak{B}_C\mathfrak{B}_u^3\big).$$

Proof. Appendix A.6.6.

Proposition 8.13 (Unstable spectral abscissa lower bound; Honarpisheh et al. [19], Theorem 4.1)**.** *Let $s_A > 0$ be fixed and let $\mathcal{S} = \{u_{(i)}\}_{i=1}^m \subset [-1,1]^T$. If $|w| \le \mathfrak{B}_w$, then*

$$\mathrm{Rad}_{\mathcal{S}}(\mathcal{F}_{\mathrm{SSM}}) \ge \mathfrak{B}_w \frac{(1+s_A)^T - 1}{s_A}\sqrt{\frac{2}{\pi m}}.$$

If $s_A = 0$, then

$$\mathrm{Rad}_{\mathcal{S}}(\mathcal{F}_{\mathrm{SSM}}) \ge \mathfrak{B}_w T\sqrt{\frac{2}{\pi m}}.$$

Proof. Appendix A.6.6.

8.5 Lyapunov Stability of MambaBlock Perturbations

Halloran et al. [18] study the Mamba and Mamba-2 SSM layer as the time-varying recurrence

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \qquad y_t = C_t h_t, \qquad \bar{A}_t = \exp(\bar{\Delta}_t A).$$

The following statement makes explicit the diagonal stability condition used by the proof.

Theorem 8.14 (Non-positive MambaBlock Lyapunov exponent; Halloran et al. [18], Theorem 1)**.** *Let*

$$A = \mathrm{diag}(a_1, \dots, a_N), \qquad \bar{\Delta}_t = \mathrm{diag}(\delta_{t,1}, \dots, \delta_{t,N}),$$

where $\delta_{t,i} \ge 0$ and $\mathrm{Re}\, a_i \le 0$. Let F_θ denote the source MambaBlock state map $h_t = F_\theta(h_{t-1}, x_t)$. Then

$$\lambda_{\max} := \limsup_{L\to\infty} \frac{1}{L}\log\left\|\prod_{t=1}^L \frac{\partial h_t}{\partial h_{t-1}}\right\|_2 \le 0.$$

Consequently, for perturbations of size ε at a latent/input state,

$$\max |F_\theta^n(h_{t-1}, x_t) - F_\theta^n(h_{t-1} + \varepsilon, x_t + \varepsilon)| \in O\big(|\varepsilon| e^{n\zeta}\big)$$

for some $\zeta \le 0$.

Proof. Appendix A.6.7.

Corollary 8.15 (Output and fine-tuning perturbations; Halloran et al. [18], Theorems 2–4)**.** *Assume the preceding MambaBlock hypotheses and $\sup_t \|C_t\|_2 < \infty$. Let $\{h_t, x_t, y_t\}$ be the full-precision trajectory and $\{h'_t, x'_t, y'_t\}$ be the trajectory under the source MPFT/PEFT perturbation model. Define*

$$\varepsilon^* = \max_{0 \le t \le L} \max\{\|h_t - h'_t\|_\infty, \|x_t - x'_t\|_\infty\}, \qquad x_0 = 0.$$

Then

$$\max \left|F_\theta^L(h_0, x_1) - F_\theta^L(h'_0, x'_1)\right| \in O\big(\varepsilon^* e^{L\zeta}\big)\,, \qquad \max |y_L - y'_L| \in O\big(\varepsilon^* e^{L\zeta}\big)\,,$$

for some $\zeta \le 0$.

Proof. Appendix A.6.7.

8.6 Training Dynamics for a Simplified Mamba Block

Shandirasegaran et al. [37] analyze a single-layer, single-head representative Mamba block with input-dependent gating and a two-layer MLP readout. The following statements use the source notation: L is the sequence length, d the ambient dimension, m the model width, N the number of training samples, η the learning rate, and T_{gd} the number of gradient-descent iterations.

The data model fixes an orthonormal family $\mathcal{O} = \{o_+, o_-, o_3, \dots, o_d\}$. Each token is a Gaussian-noisy copy of one element of $\mathcal{O}$, and the training set is balanced up to $O(\sqrt{N})$ samples between positive and negative labels. In the majority-voting regime, α_r and α_c denote the average fractions of class-relevant and confusing tokens, so $\alpha_r - \alpha_c$ is the signal gap. In the locality-structured regime, the quantities $\Delta L^+_{o_+}$ and $\Delta L^-_{o_+}$ record the separation between the two o_+ tokens in positive and negative samples, respectively.

Lemma 8.16 (Gating alignment for majority-voting data; Shandirasegaran et al. [37], Lemma 4.1)**.** *Assume each entry of $W_O^{(0)}$ is drawn independently from $\mathcal{N}(0, \xi^2)$ and $w_\Delta^{(0)} = 0$. With sufficiently many training samples and iterations,*

$$\langle w_\Delta^{(T_{\mathrm{gd}})}, o_+ \rangle \ge \frac{\eta T_{\mathrm{gd}}}{8L^2} \Theta((\alpha_r L - \alpha_c L)^2),$$

$$\langle w_\Delta^{(T_{\mathrm{gd}})}, o_- \rangle \ge \frac{\eta T_{\mathrm{gd}}}{8L^2} \Theta((\alpha_r L - \alpha_c L)^2),$$

and for every $j \ge 3$,

$$\langle w_\Delta^{(T_{\mathrm{gd}})}, o_j \rangle \le \widetilde{O}(1/\operatorname{poly}(d)).$$

Proof. Appendix A.6.8.

Theorem 8.17 (Majority-voting generalization; Shandirasegaran et al. [37], Theorem 1)**.** *Suppose $m \ge d^2 \log q$ for some constant $q > 0$ and the token noise level satisfies $\tau < O(1/d)$. With probability at least $1 - N^{-d}$, if*

$$N \ge \Omega\left(\frac{L^2 d}{\eta^2 (\alpha_r - \alpha_c)^2}\right), \qquad T_{\mathrm{gd}} = \Theta\left(\frac{L^2}{\eta(\alpha_r - \alpha_c)^2}\right),$$

then the trained model has zero value for the source target-error functional,

$$f(v^{(0)}, W_O^{(T_{\mathrm{gd}})}, w_\Delta^{(T_{\mathrm{gd}})}, W_B^{(0)}, W_C^{(0)}) = 0.$$

Proof. Appendix A.6.8.

Lemma 8.18 (Gating alignment for locality-structured data; Shandirasegaran et al. [37], Lemma 4.2)**.** *Under the same initialization, with sufficiently many training samples and iterations, for every $j \geq 3$,*

$$\langle w_\Delta^{(T_{\mathrm{gd}})}, o_+ \rangle \geq -\widetilde{O}(1/\operatorname{poly}(d)), \qquad \langle w_\Delta^{(T_{\mathrm{gd}})}, o_- \rangle \geq -\widetilde{O}(1/\operatorname{poly}(d)),$$

and

$$\langle w_\Delta^{(T_{\mathrm{gd}})}, o_j \rangle \leq -\frac{\eta T_{\mathrm{gd}} c'^3}{16L} \left[(1/2)^{\Delta L_{o_+}^+ - 2} - (1/2)^{\Delta L_{o_+}^- - 2} \right] \cdot \left[(1/2)^{\Delta L_{o_+}^+} + (1/2)^{\Delta L_{o_-}^-} \right].$$

Proof. Appendix A.6.8.

Theorem 8.19 (Locality-structured generalization; Shandirasegaran et al. [37], Theorem 2)**.** *Suppose $m \geq d^2 \log q$ for some constant $q > 0$ and $\tau < O(1/d)$. With probability at least $1 - N^{-d}$, if*

$$N \geq \Omega\left(\frac{L^2 d}{\eta^2 \left[(1/2)^{\Delta L_{o_+}^+} - (1/2)^{\Delta L_{o_+}^-} \right]^2} \right)$$

and

$$T_{\mathrm{gd}} = \Theta\left(\frac{L^2}{\eta \left[(1/2)^{\Delta L_{o_+}^+} - (1/2)^{\Delta L_{o_+}^-} \right]} \right),$$

then

$$f(v^{(0)}, W_O^{(T_{\mathrm{gd}})}, w_\Delta^{(T_{\mathrm{gd}})}, W_B^{(0)}, W_C^{(0)}) = 0.$$

Proof. Appendix A.6.8.

9 In-Context Learning and Algorithmic Behavior

The in-context-learning results below are task-model theorems. They identify particular algorithms that SSM or Mamba-like mechanisms can implement or learn: gradient descent, Laplace smoothing, online gradient descent, outlier-filtered classification, and low-dimensional feature learning. Here N is often the prompt length or number of in-context examples, following the source papers; hidden dimension is written d_h when both quantities appear in the same statement. Symbols such as T, T_1, and T_2 denote training iterations or pretraining sample counts only inside the theorem where they are defined.

9.1 SSMs as Gradient Accumulators

Proposition 9.1 (One-step gradient descent by a diagonal recurrent SSM; Sushma et al. [39], Proposition 4.1)**.** *Consider the zero-initialized linear-regression construction in Sushma et al. [39] and let*

$$\Delta w = \eta \nabla_w \mathcal{L}(D_{1:t}; w_0).$$

There is a diagonal linear recurrent layer with linear input and output gating, token-dependent recurrent matrix $A(c_j)$, input map $B(c_j)$, and output map $U(c_j)$ such that, for tokens

$$s_j = c_j = [y_j x_j, x_{j+1}], \qquad j = 1, \ldots, N,$$

each recurrent step outputs

$$\widehat{y}_{j+1} = -(\Delta w)^\top x_{j+1}.$$

The final token contains x_{N+1} and gives the test prediction $\widehat{y}_{N+1}$.

Proof. Appendix A.7.1.

Proposition 9.2 (Vector-valued, multi-step, and nonlinear variants; Sushma et al. [39], Proposition 4.2 and Propositions A.1–A.2)**.** *For vector-valued linear regression with sequence tokens $s_{2j} = x_j$ and $s_{2j+1} = y_j$, a diagonal linear recurrent layer with a sliding window of size 3 and stride 2 can be constructed so that*

$$\widehat{y}_{j+1} = -(\Delta W)^\top x_{j+1}, \qquad \Delta W = \eta \nabla_W \mathcal{L}.$$

With ℓ recurrent layers, one can construct layerwise pairs of recurrent matrices, input maps, and output maps so that the output corresponds to ℓ steps of gradient descent. If the recurrent layer is followed by an MLP, the same construction extends to a nonlinear regression model by using the MLP to map the accumulated linear statistic to the nonlinear gradient statistic.

Proof. Appendix A.7.1.

9.2 Markov Chains and Laplace Smoothing

Theorem 9.3 (Bayes-optimal add-β estimator for Markov prompts; Bondaschi et al. [2], Theorem 3)**.** *Let $P = (P_{i_1^k})_{i_1^k \in \mathcal{X}^k}$ be a k-th order binary Markov kernel, with each conditional distribution independently drawn from $\mathrm{Dir}(\beta)$. If $x_1^k \sim \mathrm{Unif}(\mathcal{X}^k)$ is independent of P, then the predictor minimizing the Bayesian cross-entropy loss is the add-β estimator*

$$\mathbb{P}_\beta^{(k)}(x_{t+1} = j \mid x_1^t) = \frac{n_j + \beta}{n + 2\beta},$$

where n_j is the number of previous occurrences of symbol j after the current length-k context and $n = n_0 + n_1$.

Proof. Appendix A.7.2.

Theorem 9.4 (MambaZero represents order-1 Laplace smoothing; Bondaschi et al. [2], Theorem 1)**.** *For the canonical MambaZero model with dimensions $d = N = 2$, expansion $e = 1$, and convolution window $w = 2$, for every $\beta > 0$ and $\varepsilon \in (0, 1)$ there are parameters θ such that, for all binary sequences $(x_t)_{t \geq 1}$ and all $t \geq 1$,*

$$\mathrm{KL}\Big(\mathbb{P}_\beta^{(1)}(\cdot \mid x_1^t) \,\Big\|\, \mathbb{P}_\theta(\cdot \mid x_1^t)\Big) \leq \varepsilon.$$

Proof. Appendix A.7.2.

Theorem 9.5 (Finite-precision recurrent lower bound; Bondaschi et al. [2], Theorem 2)**.** *Consider any recurrent predictor*

$$r_t = h(r_{t-1}, x_t), \qquad \widehat{p}_t = g(r_t), \qquad r_t \in \mathbb{R}^d,$$

with bit precision $\mathtt{p}$. Suppose $P \sim \mathrm{Dir}(1)$ and the predictor satisfies, for all sufficiently large t and almost surely over P and $x_1^t \sim P$,

$$\left\| \widehat{p}_t - \mathbb{P}_1^{(k)}(\cdot \mid x_1^t) \right\|_\infty \leq \varepsilon.$$

Then

$$d \cdot \mathtt{p} \geq 2^k (1 - 3\varepsilon) \log(1/\varepsilon).$$

The same lower bound holds for time-varying maps $r_t = h_t(r_{t-1}, x_t)$ and $\widehat{p}_t = g_t(r_t)$.

Proof. Appendix A.7.3.

9.3 Trained Mamba as Online Gradient Descent

Theorem 9.6 (Online-gradient-descent form of trained Mamba; Jiang et al. [23], Theorem 1)**.** *Assume the simplified Mamba linear-regression setting of Jiang et al. [23]: $A = -I_{d_h}$, $w_\Delta = 0$, fixed $b_\Delta = \log(\exp((\log 2)/N) - 1)$, Gaussian initialization of W_B, W_C, $d_h = \widetilde{\Omega}(d^2)$, learning rate $\eta = O(d^{-2} d_h^{-1})$, zero bias initialization, and $N = \Omega(d)$. If Mamba is trained by gradient descent, then with probability at least $1 - \delta$, as training time $t \to \infty$ the trainable parameters*

$$\theta'(t) = \{W_B(t), W_C(t), b_B(t), b_C(t)\}$$

converge to parameters satisfying

$$(W_C^\top)_{[1:d,:]} h_l^{(d+1)} = \alpha (W_C^\top)_{[1:d,:]} h_{l-1}^{(d+1)} + (1 - \alpha)\beta y_l x_l,$$

$$\widehat{y}_q = x_q^\top \sum_{i=0}^{N-1} (1-\alpha)\alpha^{i+1}\beta\, y_{N-i} x_{N-i},$$

and

$$\mathcal{L}(\theta(t)) \le \frac{3d(d+1)}{2N}.$$

Here

$$\alpha = \exp(-(\log 2)/N), \qquad \beta = \frac{2(1+\alpha)}{\alpha(3(1-\alpha)d + 4 - 2\alpha)}.$$

Proof. Appendix A.7.4.

9.4 Outliers and Gating

Theorem 9.7 (Mamba training with outlier prompts; Li et al. [27], Theorem 1)**.** *Let $p_a \in [0,1)$ be the fraction of outlier-containing context examples in training prompts and define*

$$B_T = \max\{\varepsilon^{-2},\, M_1(1-p_a)^{-1}\}\log\varepsilon^{-1},$$

$$B_M = \max\{B_T,\, \beta^{-4}V^2\kappa_a^{-2}(1-p_a)^{-2}\log\varepsilon^{-1}\}.$$

For the one-layer Mamba model, orthogonal pattern task distribution, balanced training labels, and training-task coverage condition of Li et al. [27], if

$$B \gtrsim B_M, \qquad V\beta^{-4} \lesssim \kappa_a \lesssim V\beta(1-p_a)p_a^{-1}\varepsilon^{-1},$$

$$p_a^{-1}\,\mathrm{poly}(M_1^{\kappa_a}) \gtrsim l_{\mathrm{tr}} \gtrsim (1-p_a)^{-1}\log M_1,$$

and after

$$T \ge T_M = \Theta(\eta^{-1}(1-p_a)^{-1}\beta^{-2}M_1)$$

iterations with $\eta \le 1$ and $N = BT$ samples, then

$$\mathbb{E}_{f\in\mathcal{T},\,P\sim\mathcal{D}}[\ell(\Psi^{(T)};P,z)] \le \varepsilon.$$

Proof. Appendix A.7.5.

Theorem 9.8 (Generalization to shifted outliers; Li et al. [27], Theorem 2)**.** *Let the test outlier directions belong to*

$$\mathcal{V}' = \left\{ v :\, v = \sum_{i=1}^{V}\lambda_i v_i^*,\, \sum_{i=1}^{V}\lambda_i \ge L > 0,\, \|v\| = 1 \right\}.$$

If

$$\kappa_a' \in \left[\kappa_a,\, \Theta(V\beta p_a^{-1}\kappa_a^{-1}L^{-1}(1-p_a)\varepsilon^{-1})\right],$$

$$\alpha < \min(1, p_a l_{\mathrm{tr}}/l_{\mathrm{ts}}), \qquad \alpha^{-1}\,\mathrm{poly}(M_1^{\kappa_a}) \gtrsim l_{\mathrm{ts}} \gtrsim (1-\alpha)^{-1}\log M_1,$$

then for testing prompts $P' \sim \mathcal{D}'$,

$$L^{0\text{-}1}_{f\in\mathcal{T},\,P'\sim\mathcal{D}'}(\Psi^{(T)};P',z) \le \varepsilon.$$

Proof. Appendix A.7.5.

Corollary 9.9 (Attention selection and nonlinear gating; Li et al. [27], Corollaries 1 and 2)**.** *For the model trained as above, let $\mathcal{N}_1$ be the indices of context examples sharing the query's relevant pattern. With high probability,*

$$\sum_{i\in\mathcal{N}_1} \widetilde{p}_i^\top (W_B^{(T)})^\top W_C^{(T)} \widetilde{p}_{\mathrm{query}} \ge \Theta(1),$$

whereas

$$\sum_{i\in[l_{\mathrm{ts}}]\setminus\mathcal{N}_1} \widetilde{p}_i^\top (W_B^{(T)})^\top W_C^{(T)} \widetilde{p}_{\mathrm{query}} \le \Theta((1-p_a)^{-1}\varepsilon).$$

Moreover, for an outlier-containing example,

$$G_{i,l_{\mathrm{ts}}+1}(w^{(T)}) \le O(\mathrm{poly}(M_1)^{-1}),$$

while if $h(j)$ is the j-th closest clean context example to the query, then

$$G_{h(j),l_{\mathrm{ts}}+1}(w^{(T)}) \ge \Theta(2^{-(j-1)}).$$

Proof. Appendix A.7.5.

9.5 Low-Dimensional Targets and Test-Time Feature Learning

Definition 9.10 (Information and generative exponents; Oh et al. [31])**.** For $h \in L^2(\mathcal{N}(0,1))$, write its Hermite expansion as

$$h(z) = \sum_{i=0}^{\infty} \frac{H(h,i)}{i!} \mathrm{He}_i(z), \qquad H(h,i) = \mathbb{E}[h(Z)\,\mathrm{He}_i(Z)].$$

The information exponent is

$$\mathrm{ie}(h) = \min\{i \ge 1 : H(h,i) \ne 0\},$$

and the generative exponent is

$$\mathrm{ge}(h) = \min_{\mathcal{T} \in L^2(\mathbb{P}_h)} \min\{i \ge 1 : H(\mathcal{T} \circ h, i) \ne 0\}.$$

For polynomial g_*,

$$\mathrm{ge}(g_*) = \begin{cases} 2, & g_* \text{ even}, \\ 1, & g_* \text{ not even}, \end{cases}$$

as used in Oh et al. [31].

Definition 9.11 (Low-dimensional Mamba ICL setup; Oh et al. [31])**.** The final two results use Gaussian single-index prompts. A task has an unknown feature vector β supported on an r-dimensional coordinate subspace, and examples satisfy

$$x \sim \mathcal{N}(0, I_d), \qquad y = g_*(\langle \beta, x \rangle) + \zeta.$$

Let $\phi(x)$ be the Hermite feature embedding up to degree two, and let Z be the prompt embedding of $(x_1, y_1, \ldots, x_N, y_N, x)$. Write γ for the trainable Mamba vector and γ_{init} for the scalar initialization scale. The simplified Mamba layer used in the theorem has

$$W_B^\top W_C = \mathrm{diag}(\gamma, 0), \qquad w = (0, \ldots, 0, \rho^{-1}),$$

so that its last output coordinate is

$$\mathrm{Mamba}(Z;\gamma)[\widetilde{d}+1, N+1] = \sum_{j=1}^{N} G_{j,N+1}(Z) y_j \phi(x_j)^\top (\gamma \odot \phi(x)),$$

where

$$G_{j,l}(Z) = \sigma(w^\top z_j + b) \prod_{k=j+1}^{l} \big(1 - \sigma(w^\top z_k + b)\big),$$

with z_k denoting the kth prompt-embedding column. The prediction model is the two-layer ReLU readout

$$f(Z;\gamma,u,v,a) = \sum_{k=1}^{m} u_k \,\mathrm{ReLU}\Big(v_k N^{-1}\,\mathrm{Mamba}(Z;\gamma)[\widetilde{d}+1, N+1] + a_k\Big).$$

For $T_{\mathrm{pt}} = T_1 + T_2$ pretraining tasks of context length N_{pt}, define squared losses L_1, L_2 on the first T_1 and last T_2 tasks by

$$L_\ell(\gamma,u,v,a) = \frac{1}{T_\ell} \sum_{t=T_{\ell-1}+1}^{T_{\ell-1}+T_\ell} \big(f(Z^t;\gamma,u,v,a) - y^t\big)^2, \qquad \ell = 1,2, \quad T_0 = 0.$$

Stage I initializes

$$\gamma^{(0)} = (\gamma_{\text{init}}^2, 1, \dots, 1, \gamma_{\text{init}}, \dots, \gamma_{\text{init}}), \qquad u^{(0)} = m^{-1}\mathbf{1}_m, \quad v^{(0)} = \mathbf{1}_m, \quad a^{(0)} = 0,$$

and takes one regularized gradient step

$$\gamma^* = \gamma^{(0)} - \eta \, \nabla_\gamma \left(L_1(\gamma, u^{(0)}, v^{(0)}, a^{(0)}) + \frac{\lambda_1}{2}\|\gamma\|_2^2 \right)\Big|_{\gamma=\gamma^{(0)}}.$$

Stage II fixes γ^*, samples $v^* \sim \text{Unif}(\{\pm 1\}^m)$ and $a^* \sim \text{Unif}([-1,1]^m)$, and solves the ridge problem

$$u^* = \arg\min_u \left(L_2(\gamma^*, u, v^*, a^*) + \frac{\lambda_2}{2}\|u\|_2^2 \right).$$

Proposition 9.12 (Test-time feature learned by the Mamba layer; Oh et al. [31], Proposition B.8)**.** *Let $(x_1, y_1, \dots, x_N, y_N, x)$ be a prompt from a Gaussian single-index model with feature vector $\beta \in \mathbb{R}^d$, context length $N \le N^*$, and embedding Z. If $N = \widetilde{\Omega}(r^{3\,\text{ge}(g_*)})$, the Stage-I updated parameter γ^* is the one-step gradient update above, and $\eta = \Theta(d^{2C_b}(\log d)^{-C_\eta})$ for a sufficiently large constant C_η, then with high probability*

$$N^{-1}\,\text{Mamba}(Z;\gamma^*)[\widetilde{d}+1, N+1] = P_1 + P_2\left(\frac{\langle\beta, x\rangle}{r}\right)^{\text{ge}(g_*)} + o\Big(P_2 r^{-3\,\text{ge}(g_*)/2}(\log d)^{-2\deg(g_*)+2}\Big),$$

where $P_1 = o(1)$ and $P_2 = \Theta((\log d)^{-C_{P_2}})$ are independent of the data.

Proof. Appendix A.7.6.

Theorem 9.13 (Low-dimensional single-index ICL; Oh et al. [31], Theorem 3.2)**.** *Let $f(\cdot;\gamma^*, u^*, v^*, a^*)$ be the Mamba model pretrained by the two-stage procedure above. Assume $N^*, T^* \le d^{C^*}$ for a sufficiently large constant $C^* > 0$, and suppose*

$$N_{\text{pt}} = \widetilde{\Omega}\Big(r^{3\,\text{ge}(g_*)}d^8 \vee T_1^2 d^4\Big),$$

$$T_1 = \widetilde{\Omega}(r^{3\,\text{ge}(g_*)}d^6), \qquad T_2 = \widetilde{\Omega}(r^{3\,\text{ge}(g_*)}),$$

$$m = \widetilde{\Omega}(r^{4\,\text{ge}(g_*)}),$$

with fixed weights

$$\rho = \Theta((\log d)^{C_\rho}), \qquad b = C_b \log d, \qquad \gamma_{\text{init}} = \Theta((\log d)^{-C_\gamma})$$

for sufficiently large constants $C_\rho, C_b, C_\gamma > 0$. Then there exist $\lambda_1, \lambda_2, \eta$ such that, with probability at least 0.99 over pretraining data and random initialization, if

$$N_{\text{test}} = \widetilde{\Omega}(r^{3\,\text{ge}(g_*)}),$$

then

$$\mathcal{R}_{N_{\text{test}}}(\gamma^*, u^*, v^*, a^*) \le \tau + o(1).$$

Proof. Appendix A.7.6.

A Proofs and Derivations

This appendix collects longer proofs, derivations, and notation translations used in the main text. A proof heading inherits the provenance of the corresponding main-text result: source results follow the cited theorem or section, reformulated results are algebraic restatements of displayed source formulas, and derived results are consequences proved inside this note.

A.1 Proofs: Memory and Structured SSMs

This subsection gives the derivations used by Section 2. For algorithmic results, the proof records the mathematical reduction and the structured primitive being invoked.

A.1.1 Theorem 3.2: HiPPO-LegS dynamics

This follows [11, Theorem 2 and Appendix D.3].

Proof. Let

$$p_n^{(t)}(s) = \sqrt{2n+1}\, P_n(2s/t - 1),$$

where P_n is the usual Legendre polynomial on $[-1, 1]$. The basis $\{p_n^{(t)}\}$ is orthonormal for $d\mu^{(t)}(s) = t^{-1}\mathbf{1}_{[0,t]}(s)\, ds$, and the LegS coefficient is

$$h_n(t) = \frac{1}{t}\int_0^t u(s) p_n^{(t)}(s)\, ds.$$

Differentiate this coefficient. The moving upper endpoint gives $t^{-1}u(t)p_n^{(t)}(t)$, and $p_n^{(t)}(t) = \sqrt{2n+1}$ because $P_n(1) = 1$. The derivative of the scaled measure and the derivative of the moving basis combine into the lower-triangular expansion

$$-\frac{1}{t}p_n^{(t)}(s) + \partial_t p_n^{(t)}(s) = -\frac{1}{t}\left((n+1)p_n^{(t)}(s) + \sum_{k<n}\sqrt{(2n+1)(2k+1)}\, p_k^{(t)}(s)\right).$$

Substituting this expansion into the differentiated coefficient formula and using

$$h_k(t) = \frac{1}{t}\int_0^t u(s) p_k^{(t)}(s)\, ds$$

gives

$$\dot{h}_n(t) = -\frac{1}{t}\sum_{k\le n} A_{nk} h_k(t) + \frac{1}{t}B_n u(t),$$

with the matrix A and vector B stated in the theorem. Forward Euler at integer times gives

$$h_{k+1} = \left(I - \frac{A}{k}\right) h_k + \frac{1}{k} B u_k.$$

□

A.1.2 Proposition 3.3: HiPPO-LegS timescale equivariance and fast step

This follows [11, Propositions 3–4 and Appendix E.1–E.2].

Proof. For timescale equivariance, write the LegS coefficient of u as

$$c_n^u(t) = \frac{1}{t}\int_0^t u(s)\phi_n(s/t)\, ds, \qquad \phi_n(r) = \sqrt{2n+1}P_n(2r-1).$$

If $v(s) = u(\alpha s)$, then the change of variables $r = \alpha s$ gives

$$c_n^v(t) = \frac{1}{t}\int_0^t u(\alpha s)\phi_n(s/t)\, ds = \frac{1}{\alpha t}\int_0^{\alpha t} u(r)\phi_n(r/(\alpha t))\, dr = c_n^u(\alpha t).$$

This proves the displayed identity for every coefficient.

For the fast step, the LegS matrix is a diagonal part plus diagonal scalings of a triangular cumulative-sum operator. Consequently multiplication by the LegS matrix is computed by cumulative sums in $O(N)$ time. A generalized bilinear-transform discretization only requires applying matrices of the form $I + \eta A$ and $(I - \eta A)^{-1}$ for scalar η. The triangular factorization reduces the inverse application to a first-order recurrence over coordinates, again computable by one cumulative pass. Thus each discretized recurrent update costs $O(N)$ operations. □

A.1.3 Proposition 3.4: HiPPO-LegS approximation

This follows [11, Proposition 6 and Appendix E.4].

Proof. Write the LegS expansion of the rescaled history $v_t(r) = u(tr)$, $r \in [0,1]$, as

$$v_t(r) = \sum_{n \geq 0} c_n(t)\phi_n(r), \qquad \phi_n(r) = \sqrt{2n+1}\, P_n(2r-1).$$

The degree-$(N-1)$ HiPPO projection keeps the first N coefficients. By orthonormality and Parseval,

$$\|v_t - \Pi_N v_t\|^2_{L^2([0,1])} = \sum_{n \geq N} |c_n(t)|^2.$$

If u is L_u-Lipschitz, then v_t is tL_u-Lipschitz. The standard Legendre coefficient estimate for Lipschitz functions gives $|c_n(t)| \lesssim tL_u/n$, so

$$\sum_{n \geq N} |c_n(t)|^2 \lesssim t^2 L_u^2 \sum_{n \geq N} n^{-2} \lesssim \frac{t^2 L_u^2}{N}.$$

Taking square roots gives $O(tL_u/\sqrt{N})$. If u has k bounded derivatives, repeated integration by parts against the Legendre differential equation gives $|c_n(t)| \lesssim t^k n^{-k}$, hence

$$\sum_{n \geq N} |c_n(t)|^2 \lesssim t^{2k} \sum_{n \geq N} n^{-2k} = O(t^{2k} N^{-2k+1}).$$

Taking square roots gives

$$\|v_t - \Pi_N v_t\|_{L^2([0,1])} = O(t^k N^{-k+1/2}).$$

□

A.1.4 Theorem 3.5: FouT finite-window approximation

This follows the theorems labelled 'thm:hippo-fourier' and 'thm:fourier-kernel-approx' in [15, Appendix D].

Proof. The generalized HiPPO construction differentiates coefficients in a moving orthogonal basis. For the finite-window Fourier choice on a fixed window, differentiating the sine and cosine basis functions only mixes neighboring Fourier coordinates through constant coefficients. Thus the $N \to \infty$ limit is a time-invariant SSM using the orthonormal basis

$$1, \quad c_n(r) = \sqrt{2}\cos(2\pi n r), \quad s_n(r) = \sqrt{2}\sin(2\pi n r), \qquad r \in [0,1].$$

It remains to record the approximation bound. Let $M = \lfloor N/2 \rfloor$ and write the Fourier coefficients of K as $a_n = \langle K, c_n \rangle$ and $b_n = \langle K, s_n \rangle$. Parseval gives

$$\|K - \widehat{K}_N\|_2^2 = \sum_{n \geq M} \left(|a_n|^2 + |b_n|^2\right).$$

If K is L_K-Lipschitz, integration by parts gives, for $n \geq 1$,

$$|a_n| = \left| -\frac{1}{2\pi n} \int_0^1 K'(r) s_n(r)\, dr \right| \leq \frac{L_K}{2\pi n},$$

and the same estimate holds for $|b_n|$. Therefore

$$\|K - \widehat{K}_N\|_2^2 \leq \sum_{n \geq M} \frac{2L_K^2}{(2\pi n)^2} \leq \frac{L_K^2}{\pi^2 (N-2)}.$$

Thus $\|K - \widehat{K}_N\|_2 \leq L_K/(\pi\sqrt{N-2})$, and the first displayed condition on N implies error at most ε.

If K has k derivatives bounded by L_K, repeat integration by parts k times. Periodicity of the sine and cosine basis removes boundary terms, and

$$|a_n|, \ |b_n| \leq \frac{L_K}{(2\pi n)^k}.$$

Hence

$$\|K - \widehat{K}_N\|_2^2 \le \sum_{n \ge M} \frac{2L_K^2}{(2\pi n)^{2k}} \le \frac{L_K^2}{\pi^{2k}(N-2)^{2k-1}}.$$

Taking square roots and solving $L_K/(\pi^k(N-2)^{k-1/2}) \le \varepsilon$ gives the second state-size condition. □

A.1.5 Proposition 3.6: LSSL convolution form

Proof. Unroll the recurrence in [14, Section 3.1]. Since $h_0 = 0$,

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t = \sum_{r=1}^{t} \bar{A}^{t-r}\bar{B}x_r.$$

Multiplying by C and adding the skip term gives

$$y_t = Dx_t + \sum_{r=1}^{t} C\bar{A}^{t-r}\bar{B}x_r = Dx_t + \sum_{j=0}^{t-1} C\bar{A}^j\bar{B}\,x_{t-j}.$$

Thus the causal convolution kernel is $K_j = C\bar{A}^j\bar{B}$, i.e. the Krylov sequence $\mathcal{K}_L(\bar{A}, \bar{B}, C)$. □

A.1.6 Lemma 3.7: Gate as discretization

Proof. Apply backward Euler to $\dot{h} = -h + x$ with step size $\Delta t > 0$:

$$\frac{h_t - h_{t-1}}{\Delta t} = -h_t + x_t.$$

Solving for h_t gives

$$h_t = \frac{1}{1+\Delta t}h_{t-1} + \frac{\Delta t}{1+\Delta t}x_t.$$

Set $\Delta t = \exp(z)$. Then

$$\frac{\Delta t}{1+\Delta t} = \frac{e^z}{1+e^z} = \sigma(z), \qquad \frac{1}{1+\Delta t} = 1 - \sigma(z),$$

which is exactly the gated recurrence. □

A.1.7 Theorem 3.8: Fast Krylov computation

Proof. The generating function of the Krylov sequence is

$$\sum_{j \ge 0} CA^jB\,z^j = C(I - zA)^{-1}B.$$

Thus computing $\mathcal{K}_L(A, B, C)$ is equivalent to computing the first L coefficients of this rational function. For a k-quasiseparable matrix with constant k, [14, Appendix E.1] represents the relevant resolvent entries by a recursive product of low-rank generators. Each merge step is polynomial multiplication and truncation, so divide-and-conquer with fast polynomial arithmetic evaluates the first L coefficients in $\widetilde{O}(N+L)$ time and space and logarithmic depth in the structured-matrix model. This proves the theorem after identifying the coefficients with the convolution kernel. □

A.1.8 Lemma 3.9: Conjugation invariance

Proof. Let $h = V\widetilde{h}$. Then

$$V\dot{\widetilde{h}} = AV\widetilde{h} + Bu,$$

so

$$\dot{\widetilde{h}} = V^{-1}AV\widetilde{h} + V^{-1}Bu.$$

The output is

$$y = Ch = CV\widetilde{h}.$$

Thus the transformed triple $(V^{-1}AV, V^{-1}B, CV)$ produces exactly the same input-output map. □

A.1.9 Lemma 3.10: HiPPO diagonalization

Proof. [13, Lemma 3.2 and Appendix B.1] gives an explicit eigenvector matrix for the HiPPO matrix,

$$V_{ij} = \binom{i+j}{i-j}.$$

Substituting this formula into the eigenvector relation verifies $AV = V\Lambda$ column by column, using the standard binomial recurrence $\binom{m}{r} = \binom{m-1}{r} + \binom{m-1}{r-1}$. The conditioning statement comes from inspecting large entries of V. Along the relevant diagonal one finds binomial coefficients of order

$$\binom{4i}{2i} \sim \frac{2^{4i}}{\sqrt{2\pi i}},$$

by Stirling's formula. Taking i proportional to N gives entries of size about $2^{4N/3}$, so direct diagonalization is numerically ill-conditioned. □

A.1.10 Theorem 3.11: NPLR HiPPO representation

Proof. The proof in [13, Theorem 1 and Appendix C.1] constructs low-rank matrices P, Q such that

$$N := A + PQ^\top$$

is normal. By the spectral theorem, $N = V\Lambda V^*$ for a unitary V and diagonal Λ. Therefore

$$A = V\Lambda V^* - PQ^\top.$$

Conjugating by V gives

$$V^*AV = \Lambda - (V^*P)(V^\top Q)^\top,$$

which is diagonal plus low rank. For the HiPPO variants used in S4, the correction is written explicitly and normality is checked by direct matrix algebra; the rank is 1 or 2 depending on the variant. □

A.1.11 Theorem 3.12: S4 complexity

Proof. Write the state matrix in DPLR form $A = \Lambda - PQ^*$ after the unitary change of basis. Bilinear discretization requires applying matrices of the form

$$(I - \tfrac{\Delta}{2}A)^{-1}(I + \tfrac{\Delta}{2}A) \quad \text{and} \quad (I - \tfrac{\Delta}{2}A)^{-1}B.$$

Since $I - \frac{\Delta}{2}A$ is diagonal plus low rank, Woodbury gives

$$(D + UV^*)^{-1} = D^{-1} - D^{-1}U(I + V^*D^{-1}U)^{-1}V^*D^{-1}.$$

For constant rank, this costs $O(N)$ per recurrent step.

For convolution, let

$$\widehat{K}(z) = C(I - z\bar{A})^{-1}\bar{B}$$

be the truncated generating function. Evaluating $\widehat{K}$ at the Lth roots of unity and applying an inverse FFT recovers the length-L kernel. The same Woodbury identity reduces each resolvent evaluation to a constant number of Cauchy matrix-vector products of the form

$$\sum_n \frac{a_n}{z - \lambda_n}.$$

The Cauchy primitive used by [13, Appendix C] evaluates these products in quasi-linear time, yielding the stated $\widetilde{O}(N + L)$ convolution complexity. □

A.1.12 Proposition 3.13: DSS diagonal kernel

Proof. Diagonalize $A = V\Lambda V^{-1}$. The zero-order-hold discretization has

$$\bar{A} = e^{A\Delta} = Ve^{\Lambda\Delta}V^{-1}, \qquad \bar{B} = A^{-1}(e^{A\Delta} - I)B.$$

The kernel entries are

$$K_k = C\bar{A}^k\bar{B} = CVe^{\Lambda k\Delta}V^{-1}A^{-1}(e^{A\Delta} - I)B.$$

Absorb CV, $V^{-1}B$, and the diagonal factors into weights $\widetilde{w}_i$ to obtain

$$K_k = \sum_i \widetilde{w}_i\lambda_i^{-1}(e^{\lambda_i\Delta} - 1)e^{\lambda_i k\Delta}.$$

This is the first displayed form. For the finite horizon $0 \le k < L$, multiply and divide each term by $e^{L\lambda_i\Delta} - 1$; this rewrites the same exponential family as a row-softmax over $P_{ik} = \lambda_i k\Delta$, with renormalized weights w_i. □

A.1.13 Proposition 3.14: S4D Vandermonde kernel

Proof. If $\bar{A} = \mathrm{diag}(\bar{A}_1, \dots, \bar{A}_N)$, then

$$\bar{A}^\ell = \mathrm{diag}(\bar{A}_1^\ell, \dots, \bar{A}_N^\ell).$$

Thus

$$\bar{K}_\ell = C\bar{A}^\ell\bar{B} = \sum_{n=1}^{N} C_n\bar{B}_n\bar{A}_n^\ell.$$

Stacking the entries $\ell = 0, \dots, L-1$ gives

$$\bar{K} = (\bar{B}^\top \circ C)\mathcal{V}_L(\bar{A}),$$

where $\mathcal{V}_L(\bar{A})_{n\ell} = \bar{A}_n^\ell$. □

A.1.14 Theorem 3.15: S4D-LegS limit

Proof. The S4D-LegS normal approximation removes a rank-one term from the finite HiPPO-LegS operator. Let $\{\phi_n\}_{n<N}$ be the corresponding orthonormal system under the exponentially warped measure. The removed rank-one term is a projection onto a function represented by this same basis. As $N \to \infty$, completeness of the LegS basis gives convergence of the finite projection of that function to the full rank-one correction in L^2.

Therefore the basis generated by the normal part $A^{(N)}$ together with the scaled input vector $B/2$ converges to the basis generated by the original HiPPO-LegS pair (A, B). Since $A^{(N)}$ is normal, it is unitarily diagonalizable. Passing to this diagonal basis yields the S4D-LegS initialization and preserves the limiting kernel. □

A.1.15 Proposition 3.16: DLR equivalence

Proof. Unrolling the original linear RNN gives the convolution kernel

$$K_j = CA^jB.$$

If $A = V\Lambda V^{-1}$, then

$$K_j = CV\Lambda^jV^{-1}B = \sum_i (CV)_i(V^{-1}B)_i\lambda_i^j.$$

Set $w_i = (CV)_i(V^{-1}B)_i$. The diagonal RNN with transition Λ and readout weights w has kernel $\langle w, \Lambda^j\rangle$, hence the same input-output map. □

A.1.16 Claim 3.17: Real-positive DLR lower bound

Proof. To represent a one-hot shift kernel, the coefficients w_i must solve a Vandermonde system

$$\sum_i w_i\lambda_i^j = \mathbf{1}\{j = N\}, \qquad j = 0, \dots, N.$$

For $\lambda_i \in [0,1]$, the inverse Vandermonde formula expresses some coefficient w_i as a ratio whose denominator contains products $\prod_{r\neq i} |\lambda_i - \lambda_r|$. The maximal separation of $N+1$ points in $[0,1]$ is controlled by the Fekete determinant bound. Applying this bound to the inverse system gives

$$\max_i |w_i| \geq 2^{2N - O(\log N)}.$$

Thus real-positive diagonal recurrences can represent the shift only with exponentially large weights. □

A.1.17 Proposition 3.19: finite-horizon diagonal universality

This is Proposition `result:uni` of Ran-Milo et al. [34].

Proof. Let $k_0, \ldots, k_{t-1}$ be the first t entries of the impulse response of ϕ. Choose $N \geq t$ distinct scalars $a_1, \ldots, a_N \in \mathbb{K}$, and form the $t \times N$ Vandermonde matrix

$$V_{\ell n} = a_n^\ell, \qquad 0 \leq \ell < t, \quad 1 \leq n \leq N.$$

The first t columns with distinct a_n have full rank t. Therefore there is $w \in \mathbb{K}^N$ such that

$$Vw = (k_0, \ldots, k_{t-1})^\top.$$

Set $A = \mathrm{diag}(a_1, \ldots, a_N)$, $B = w$, and $C = \mathbf{1}$. Then

$$C^\top A^\ell B = \sum_{n=1}^{N} w_n a_n^\ell = k_\ell, \qquad 0 \leq \ell < t.$$

This gives the required finite-horizon match. The same Vandermonde argument works over $\mathbb{R}$ and over $\mathbb{C}$. □

A.1.18 Proposition 3.20: complex contains real

This is Proposition `result:c4r` of Ran-Milo et al. [34].

Proof. If $N_\mathbb{C} = N_\mathbb{R}$, take

$$A_\mathbb{C} = A_\mathbb{R}, \qquad B_\mathbb{C} = B_\mathbb{R}, \qquad C_\mathbb{C} = C_\mathbb{R}$$

and regard the real entries as complex entries with zero imaginary part. The impulse responses agree term by term:

$$C_\mathbb{C}^\top A_\mathbb{C}^\ell B_\mathbb{C} = C_\mathbb{R}^\top A_\mathbb{R}^\ell B_\mathbb{R}.$$

If $N_\mathbb{C} > N_\mathbb{R}$, append $N_\mathbb{C} - N_\mathbb{R}$ unused coordinates by setting the corresponding entries of $B_\mathbb{C}$ or $C_\mathbb{C}$ to zero. These modes contribute nothing to the impulse response. □

A.1.19 Theorem 3.21: one complex oscillator

This follows Theorem `result:r4c` of Ran-Milo et al. [34].

Proof. The complex one-mode impulse response is

$$k_\ell = B_\mathbb{C} C_\mathbb{C} A_\mathbb{C}^\ell.$$

The assumptions $|\sin(\arg A_\mathbb{C})| \geq 0.2$ and $|A_\mathbb{C}| \geq 0.5^{1/t}$ imply that along the first t time steps, the real-valued output sequence extracted from this complex mode has $\Theta(t)$ robust oscillations: on the odd-indexed subsequence it alternates between values at least $1/4$ and at most $-1/4$ after the source's normalization.

A real diagonal SSM impulse response is a real exponential sum

$$\widehat{k}_\ell = \sum_{n=1}^{N_\mathbb{R}} w_n a_n^\ell, \qquad w_n = C_{\mathbb{R},n} B_{\mathbb{R},n}.$$

Such a sequence has limited sign oscillation. Split the terms according to the sign of a_n and then restrict to an even or odd subsequence. On either subsequence each term has the form $\widetilde{w}_n (a_n^2)^m$ with $a_n^2 \geq 0$. After merging

equal bases, the functions $m \mapsto b_n^m$ with distinct $b_n \geq 0$ form a discrete Chebyshev system: if $\sum_{n=1}^{r} \widetilde{w}_n b_n^m$ has r distinct nonnegative bases, its number of sign changes is at most $r-1$. This follows by applying the finite-difference operator that annihilates one base,

$$(\mathcal{D}_b s)_m = s_{m+1} - b s_m,$$

and inducting on r. Hence an $N_{\mathbb{R}}$-term real diagonal SSM has only $O(N_{\mathbb{R}})$ robust sign alternations across the two parity subsequences. The complex target has $\Theta(t)$ alternating intervals separated by the $1/4$ margin, so the source's margin count gives that approximation within ϵ is possible only if

$$N_{\mathbb{R}} \geq \lfloor t/9 \rfloor - 1 - 4\epsilon.$$

This is the stated lower bound. □

A.1.20 Theorem 3.22: forward-difference lower bound

This is Theorem `result:r4general` of Ran-Milo et al. [34].

Proof. For a real diagonal SSM, the odd and even restrictions of the impulse response are still sums of at most $N_{\mathbb{R}}$ real exponentials:

$$\widehat{k}_{\sigma,m} = \sum_{n=1}^{N_{\mathbb{R}}} w_n b_{n,\sigma}^m, \qquad |w_n| \leq \|C_{\mathbb{R}}^{\top} \odot B_{\mathbb{R}}\|_{\infty}.$$

For a scalar exponential sequence b^m, the dth forward difference equals

$$\Delta^d(b^m) = b^m (b-1)^d.$$

For the odd/even restrictions used in the theorem, the rescaled real bases $b_{n,\sigma}$ lie in $[0,1]$. For $0 \leq b \leq 1$, the function $b^m(1-b)^d$ is maximized at $b = m/(m+d)$, and hence

$$|b_{n,\sigma}|^m |b_{n,\sigma} - 1|^d \leq \left(\frac{m}{m+d}\right)^m \left(\frac{d}{m+d}\right)^d.$$

The elementary bound

$$\left(\frac{m}{m+d}\right)^m \left(\frac{d}{m+d}\right)^d \leq 2^{-2\min\{d,m\}}$$

is Lemma `bounds_on_multiplication_of_parts` in the source. Thus bases close to 1 pay through $|b-1|^d$, while bases away from 1 pay through their m-step decay. Therefore each scalar mode contributes at most $|w_n| 2^{-2\min\{d,m\}}$ to the dth difference. By linearity,

$$\left| \left(\widehat{k}|_\sigma\right)_m^{(d)} \right| \leq N_{\mathbb{R}} \|C_{\mathbb{R}}^{\top} \odot B_{\mathbb{R}}\|_{\infty} 2^{-2\min\{d,m\}}.$$

If $\widehat{k}$ ϵ-approximates the target impulse response k up to time t, then

$$\Delta^d(k - \widehat{k})_m = \sum_{r=0}^{d} (-1)^{d-r} \binom{d}{r} (k - \widehat{k})_{m+r},$$

so

$$|\Delta^d(k - \widehat{k})_m| \leq \epsilon \sum_{r=0}^{d} \binom{d}{r} = 2^d \epsilon.$$

Thus

$$|(\widehat{k}|_\sigma)_m^{(d)}| \geq |(k|_\sigma)_m^{(d)}| - 2^d \epsilon.$$

Combining this lower bound with the real-exponential upper bound and rearranging gives, for every admissible d, m, σ,

$$N_{\mathbb{R}} \|C_{\mathbb{R}}^{\top} \odot B_{\mathbb{R}}\|_{\infty} \geq 2^{d+2\min\{d,m\}} \left(2^{-d} |(k|_\sigma)_m^{(d)}| - \epsilon\right).$$

Taking the maximum over admissible d, m, σ gives the theorem. □

A.1.21 Proposition 3.23: complex DFT construction

This is Proposition `result:c4general` of Ran-Milo et al. [34].

Proof. Let $k = (k_0, \dots, k_{t-1}) = \phi(\mathbf{I})_{:t}$. Choose t distinct complex roots of unity $\omega_n = e^{2\pi i n/t}$ and set

$$A_{\mathbb{C}} = \operatorname{diag}(\omega_0, \dots, \omega_{t-1})$$

on the first t coordinates, with any unused coordinates zeroed out if $N_{\mathbb{C}} > t$. Let F be the unitary discrete Fourier matrix, $F_{\ell n} = t^{-1/2}\omega_n^{\ell}$. The inverse DFT gives coefficients $\alpha = F^* k$ such that

$$k_\ell = \sum_{n=0}^{t-1} \alpha_n t^{-1/2} \omega_n^{\ell}.$$

Set $C_{\mathbb{C}} = t^{-1/2}\mathbf{1}$ and $B_{\mathbb{C}} = \alpha$ on these coordinates. Then

$$C_{\mathbb{C}}^{\top} A_{\mathbb{C}}^{\ell} B_{\mathbb{C}} = k_\ell, \qquad 0 \le \ell < t.$$

Since F is unitary, Plancherel gives

$$\|B_{\mathbb{C}}\|_2 = \|\alpha\|_2 = \|k\|_2 \le 2\|k\|_2, \qquad \|C_{\mathbb{C}}^{\top}\|_2 = 1,$$

where the theorem statement keeps the source's looser constant. □

A.1.22 Proposition 3.24: S5 scan and complexity

Proof. After diagonalization and zero-order-hold discretization, each S5 step has the affine form

$$h_t = \bar{\Lambda}_t h_{t-1} + \bar{B}_t x_t.$$

Affine maps compose associatively:

$$(A_2, b_2) \circ (A_1, b_1) = (A_2 A_1, A_2 b_1 + b_2).$$

Therefore all hidden states can be computed by a parallel prefix scan over the pairs $(\bar{\Lambda}_t, \bar{B}_t x_t)$. The remaining costs are the input projection into the latent MIMO state, the scan itself, and the output projection. In the paper's accounting, offline S4 evaluation has order $O(H^2 L + HL \log L)$, while S5 has order $O(PHL + PL)$. Online generation has the same order when $P = O(H)$ and the S4 state size is of order H. Hence, for $P = O(H)$, S5 matches the runtime and memory order of S4 in the source comparison. □

A.1.23 Proposition 3.25: Constrained S4/S5 relationship

Proof. Under the shared state matrix and shared timescale assumptions, each of the H S4 channels evolves with the same discretized transition $\bar{A}$ and channel-specific input vector. Stack the H channel states into a single MIMO state. Linearity gives

$$h_t^{\mathrm{S5}} = \sum_{r=1}^{H} P_r h_t^{(r)}$$

for fixed projections P_r determined by the S5 input and output matrices. Thus the constrained S5 output is a linear projection of the collection of constrained S4 latent states. The corresponding output matrix is

$$C^{\mathrm{equiv}} = [\, C \mid C \mid \cdots \mid C \,],$$

whereas the ordinary block-diagonal S4 output matrix keeps the channel readouts on separate diagonal blocks. Thus the proposition identifies equivalent dynamics and a modified output projection, not equality with the usual block-diagonal S4 outputs. □

A.1.24 Proposition 3.26: H3 associative recall

Construction. Let

$$L_K = \{k_1, k_2, k_3, k_4\}, \qquad L_V = \{v_1, v_2, v_3, v_4\}.$$

A sequence in the source language has even length N of key-value examples followed by a query key:

$$x_1, x_2, \ldots, x_N, x_{N+1}.$$

Odd positions are keys, even positions are their values $x_{2i} = f_x(x_{2i-1})$, and the query x_{N+1} is one of the previously appearing keys. The desired next token is $f_x(x_{N+1})$.

Embed keys as $e_1, \ldots, e_4$ and values as $e_5, \ldots, e_8$ in $\mathbb{R}^8$. The construction uses four H3 heads, one per key. The query and key projections are chosen so that head i is active only on token k_i:

$$W_Q = W_K = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The value projection encodes each value token by two binary coordinates and sends the same encoding to every head:

$$W_V = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The first SSM in each head is a one-step shift: with state dimension $m = 2$, take $B_{\text{shift}} = e_1$ and $C_{\text{shift}} = [0\ 1]$. Thus the key signal at time $t-1$ is available at time t. The second SSM is diagonal with $A_{\text{diag}} = I$, $B_{\text{diag}} = e_1$, and $C_{\text{diag}} = e_1^\top$, so it cumulatively sums its input.

Suppose the final query is $x_{N+1} = k_i$. Then the query vector in head i is nonzero and the query vectors in all heads $j \neq i$ are zero. By the multiplicative H3 interaction, only head i can contribute. In that head, the shifted key signal is nonzero exactly at positions t for which $x_{t-1} = k_i$. At those positions, by construction of the language, $x_t = f_x(k_i)$. Therefore the diagonal cumulative-sum SSM in head i sums copies of the same two-bit encoding of $f_x(k_i)$, while all other heads output zero. A fixed final decoder maps any positive multiple of that binary encoding to the embedding of $f_x(k_i)$. Hence the constructed H3 model solves Λ. □

A.2 Proofs: Selective SSMs and Linear CDEs

A.2.1 Theorem 4.5: Mamba bases approximate Haar wavelets

This follows the wavelet construction of Huang et al. [21, Appendix].

Proof. Write the shifted Heaviside function as $H_a(s) = \mathbf{1}_{\{s \geq a\}}$. A Haar wavelet at scale j and location k decomposes as

$$\psi_{j,k}(s) = 2^{j/2} \left(H_{a_0}(s) - 2H_{a_1}(s) + H_{a_2}(s) \right),$$

where

$$a_0 = 2^{-j}k, \qquad a_1 = 2^{-j}k + 2^{-(j+1)}, \qquad a_2 = 2^{-j}(k+1).$$

It is therefore enough to realize each $2^{j/2}H_a$ as a limiting Mamba basis.

Fix a threshold $a \in [0,1]$. Use the time-augmented signal $\widetilde{x}_s = (x_s, s)$ and let the selective step depend only on the time coordinate:

$$\Delta_w(\widetilde{x}_s) = \mathrm{softplus}(w(s-a)), \qquad \lambda = 1, \qquad B = 2^{j/2}.$$

With the sign convention used in the source, take $w \to -\infty$. Then $\Delta_w(\widetilde{x}_s) \to \infty$ for $s < a$ and $\Delta_w(\widetilde{x}_s) \to 0$ for $s > a$. Thus the basis

$$g_w^{\mathrm{M}}(s; 1, \widetilde{x}) = 2^{j/2} \exp\left(- \int_s^1 \Delta_w(\widetilde{x}_r)\, dr \right)$$

converges to $2^{j/2}H_a(s)$ in the limiting-gate model: if $s < a$, the integral crosses an interval on which the gate diverges, and the exponential vanishes; if $s > a$, the limiting gate on $[s,1]$ is zero, and the exponential equals one. At the jump, the source uses the Heaviside convention matched by the limiting construction.

Applying this construction with thresholds a_0, a_1, a_2 gives three Mamba bases $g_{a_0}^{\mathrm{M}}$, $g_{a_1}^{\mathrm{M}}$, and $g_{a_2}^{\mathrm{M}}$. Their linear combination

$$g_{a_0}^{\mathrm{M}} - 2g_{a_1}^{\mathrm{M}} + g_{a_2}^{\mathrm{M}}$$

is the desired Haar wavelet in the limit. Choosing finite gate parameters sufficiently far along the limiting sequence gives the stated ϵ-approximation in the source theorem. □

A.2.2 Corollary 4.6: adaptive wavelet approximation

This is the approximation-rate corollary of Huang et al. [21].

Proof. The adaptive Haar expansion of a piecewise-constant function with m discontinuities uses only the wavelets whose supports intersect the jump set, together with the constant term. At dyadic scale j, each discontinuity belongs to one dyadic interval, so at most $O(m)$ Haar functions have nonzero coefficients. Keeping all selected coefficients through scale J therefore uses

$$M = O(mJ)$$

terms. The remaining error is supported only on the union of the $O(m)$ dyadic intervals of length 2^{-J} containing jumps. Since the function is bounded on each side of each jump, its squared error is $O(2^{-J})$ up to a constant depending on the jump heights. Rewriting J in terms of M gives the adaptive rate

$$\left\| \rho - \sum_{\ell=1}^{M} c_\ell \psi_\ell \right\|_{L^2} = O(2^{-M/m}).$$

By Theorem 4.5, each Haar function ψ_ℓ is approximated by three Mamba basis functions. Thus N Mamba bases implement $M = N/3$ adaptive Haar terms, giving

$$\left\| \rho - \sum_{n=1}^{N} g_n^{\mathrm{M}} \right\|_{L^2} = O(2^{-(N/3)/m}) = O(2^{-N/(3m)}).$$

In the comparison made in the source, the S4D basis $B_n e^{-\lambda_n(t-s)}$ is treated as a Fourier-type smooth basis for this discontinuous target class, yielding the stated $O(N^{-1})$ rate. □

A.2.3 Lemma 4.7: sensitivity decay

This is the product-rule calculation behind Huang et al. [21, Lemma `lemma_exp_decay`].

Proof. Unrolling

$$h_t = \bar{\Lambda}_t h_{t-1} + \bar{B}_t x_t$$

gives

$$h_t = \sum_{s=1}^{t} \left(\prod_{r=s+1}^{t} \bar{\Lambda}_r \right) \bar{B}_s x_s + \left(\prod_{r=1}^{t} \bar{\Lambda}_r \right) h_0.$$

The terms depending on x_j are the direct injection $\bar{B}_j x_j$ and the transition $\bar{\Lambda}_j$, which multiplies all earlier injected terms. Differentiating the unrolled expression therefore gives

$$\begin{aligned} \frac{\partial h_t}{\partial x_j} = & \left(\prod_{r=j+1}^{t} \bar{\Lambda}_r \right) \frac{\partial}{\partial x_j} (\bar{B}_j x_j) \\ & + \left(\prod_{r=j+1}^{t} \bar{\Lambda}_r \right) \frac{\partial \bar{\Lambda}_j}{\partial x_j} \left[\sum_{s=1}^{j-1} \left(\prod_{r=s+1}^{j-1} \bar{\Lambda}_r \right) \bar{B}_s x_s \right], \end{aligned}$$

which is the displayed formula in the lemma.

For S4D, the nth diagonal transition is

$$\bar{\Lambda}_{r,n} = e^{-\lambda_n}$$

and the B coefficient is independent of the future time t. The product from $r = j+1$ to t is therefore

$$\prod_{r=j+1}^{t} \bar{\Lambda}_{r,n} = e^{-\lambda_n (t-j-1)}$$

up to the same endpoint convention as the source. All remaining factors depend only on $(\lambda_n, B_n, x_{\le j})$, so they form $\widetilde{c}_{\mathrm{S4D}}$.

For S6/Mamba, the nth transition is

$$\bar{\Lambda}_{r,n} = e^{-\lambda_n \Delta(x_r)}$$

under the diagonal discretization used in the source. Hence

$$\prod_{r=j+1}^{t} \bar{\Lambda}_{r,n} = \exp\left(-\lambda_n \sum_{r=j+1}^{t} \Delta(x_r) \right).$$

The differentiated bracket depends on the local derivative at x_j and on the history before j, but not on the terminal length t. This is $\widetilde{c}_{\mathrm{M}}(\Delta, \lambda_n, B_n, x_{\le j})$. □

A.2.4 Lemma 4.8: freezing selective time

This is the retention consequence of the previous sensitivity formula in Huang et al. [21, Lemma `lemma:constant`].

Proof. For the nth S6 coordinate, isolate the direct local term in the sensitivity formula:

$$D_j = \frac{\partial}{\partial x_j} \big(B_n(x_j) \Delta(x_j) x_j \big).$$

The source writes the same local factor with its own running index convention. The future transition product is

$$\prod_{r=j+1}^{t} e^{-\lambda_n \Delta(x_r)} = \exp\left(-\lambda_n \sum_{r=j+1}^{t} \Delta(x_r) \right).$$

If

$$\lim_{t\to\infty} \lambda_n \sum_{r=j+1}^{t} \Delta(x_r) \le c,$$

then

$$\liminf_{t\to\infty} \prod_{r=j+1}^{t} e^{-\lambda_n \Delta(x_r)} \ge e^{-c}.$$

Keeping the direct local sensitivity contribution gives

$$\liminf_{t\to\infty} \left| \frac{\partial h_t^{\mathrm{M},n}}{\partial x_j} \right| \ge e^{-c} |D_j|,$$

which is the stated lower bound. Thus a coordinate can keep sensitivity to a past input only if the accumulated selective time after that input is bounded. Equivalently, along the retained part of the sequence the factors $e^{-\lambda_n \Delta(x_r)}$ must stay close enough to one. □

A.2.5 Theorem 4.9: MQAR constructions

This collects the three constructive MQAR results of Huang et al. [21].

Proof. An MQAR instance contains κ key-value pairs

$$(k_1, v_1), \ldots, (k_\kappa, v_\kappa)$$

followed by query tokens. On a query key k_i, the desired output is the stored value v_i. The constructions are exact representability statements: they specify embeddings, short convolutions, selective storage maps, and output maps.

Mamba without gating. Start with a one-hot construction of dimension $d = \kappa + |V|$. Allocate the first κ coordinates to the keys and the remaining $|V|$ coordinates to values. A key k_i is represented by a scaled key basis vector in coordinate i, and a value v is represented in the value subspace. The size-2 nonlinear convolution is chosen so that adjacent key-value pairs survive while non-pair adjacencies are suppressed. Because the vocabulary and the list of allowed adjacent patterns are finite, this can be done by a finite-margin separator: give each allowed pair a positive score, give every forbidden pair a score below a negative bias, and choose the margin larger than the maximum within-pair perturbation. The resulting local feature is nonzero exactly on key-value adjacencies.

The S6 state is organized as a matrix

$$H_t \in \mathbb{R}^{d\times\kappa},$$

whose ith column stores the value associated with key k_i. On the position containing the pair (k_i, v_i), the input-dependent map B_t selects column i, and the injected vector is the embedding of v_i. On a query position containing k_i, the input-dependent readout C_t selects column i. Thus

$$C_t H_t = \mathrm{emb}(v_i)$$

at the query, and a fixed final decoder returns v_i.

The one-hot value subspace can be reduced by a Johnson–Lindenstrauss map. Choose a random projection that preserves inner products among the finite set of value embeddings with a constant margin. After adding the source's nonnegativity/shift trick, the same storage and decoding inequalities remain valid with value dimension $O(\log |V|)$. The key-indexing subspace remains κ-dimensional, giving

$$d = O(\kappa + \log |V|), \qquad N = \kappa.$$

Mamba-2 without gating. The Mamba-2 construction uses the extra independence between the convolutional features that feed u, B, and C. In the source notation, set the B-feature convolution to a one-step shift,

$$\mathrm{conv}_B = (1, 0),$$

and take the u and C feature convolutions to be identities. Then the storage map sees the previous key while the value stream supplies the current value, exactly aligning each pair. Because the key and value features are

now separated before the SSM storage operation, Johnson–Lindenstrauss reduction can be applied to both the key-indexing and value subspaces. This gives the source dimensions

$$d = O(\log \kappa + \log |V|), \qquad N = O(\log \kappa).$$

S4D mixer inside a Mamba block. With an S4D mixer, B and C are no longer input-dependent storage and retrieval maps. The construction compensates by arranging the hidden features along the embedding dimension: split the embedding into κ chunks, one for each key, and store the value embedding for key k_i in the ith chunk. The gated convolution/output part of the Mamba block selects the queried chunk at readout. Starting from a one-hot value representation gives $O(\kappa|V|)$ embedding dimension; applying the same finite-set Johnson–Lindenstrauss reduction to each value chunk reduces this to

$$d = O(\kappa \log |V|).$$

The S4D recurrence itself uses one scalar state coordinate in the source construction, so $N = 1$. □

A.2.6 Lemma 4.10: induction head

This is the state-dimension-selective construction of Huang et al. [21, Lemma `lem_ind_head`].

Proof. For a sequence $x_1, \ldots, x_t$ over V, the induction-head target at time t is

$$y_t = x_{j(t)+1}, \qquad j(t) = \max\{j < t : x_j = x_t\},$$

with a blank output if there is no previous occurrence. Index the state dimension by tokens in V, so $N = |V|$. Use an embedding of size $2|V|$: the first copy identifies the key role and the second copy stores the value role. The size-2 convolution transforms adjacent tokens into pair features, so that after convolution the feature at position i contains the key x_{i-1} and the value x_i.

The mixer is

$$h_t = \exp(\Lambda \odot (\mathbf{1}_d \otimes \Delta(\widehat{x}_t))) \odot h_{t-1} + \widehat{x}_t \otimes B_t.$$

Choose the selective gate so that, when the current pair has key token a, the ath state column is erased and all other columns are retained. In the source construction this is obtained by taking

$$\Delta(\widehat{x}_t) = \mathrm{softplus}\big(w_\Delta [I_{|V|}\ 0]\widehat{x}_t\big)$$

and sending the gate scale to the limiting regime. The transition factor in the column corresponding to the current key tends to zero, while the transition factors in the other columns tend to one. The injection $\widehat{x}_t \otimes B_t$ then writes the current value into the erased column.

Inductively, after processing position t, the column indexed by a token $a \in V$ stores the token that followed the latest previous occurrence of a. If the current query token is $x_t = a$, the readout vector C_t selects the ath column:

$$h_t C_t = h_t e_a.$$

The fixed output matrix keeps the value half of the embedding, producing $x_{j(t)+1}$. The state has one column per vocabulary symbol and the embedding has two vocabulary copies, so

$$N = |V|, \qquad d = 2|V|.$$

□

A.2.7 Proposition 4.12: S4 and S6 as Linear CDEs

This follows [4, Section 3].

Proof. For S4, set

$$\omega_t^X = t, \qquad \xi_t^X = \int_0^t X_s \, ds.$$

Then $d\omega_t^X = dt$ and $d\xi_t^X = X_t \, dt$. The one-channel Linear CDE

$$dZ_t^X = A Z_t^X \, d\omega_t^X + B \, d\xi_t^X$$

therefore becomes

$$dZ_t^X = AZ_t^X\,dt + BX_t\,dt,$$

which is the continuous-time S4 channel.

For the S6/Mamba channel, write the discrete update as

$$z_{i,\ell} = \bar{A}_i(x_\ell)z_{i,\ell-1} + \bar{B}_i(x_\ell)x_\ell^i, \qquad \bar{A}_i(x) = \exp(\Delta_i(x)A_i),$$

with

$$\Delta_i(x) = \mathrm{softplus}(\alpha_i \cdot x + \beta_i).$$

Introduce a mesh size $\delta > 0$ and write

$$\alpha_i = \delta\widetilde{\alpha}_i, \qquad \beta_i = \delta\widetilde{\beta}_i.$$

Approximating softplus by $\sigma = \mathrm{ReLU}$ in the small-step scaling gives

$$\Delta_i(x) \simeq \delta\sigma(\widetilde{\alpha}_i \cdot x + \widetilde{\beta}_i).$$

Hence

$$\begin{aligned}\bar{A}_i(x) &= \exp\Big(\delta A_i\sigma(\widetilde{\alpha}_i \cdot x + \widetilde{\beta}_i)\Big)\\ &= I + \delta A_i\sigma(\widetilde{\alpha}_i \cdot x + \widetilde{\beta}_i) + o(\delta).\end{aligned}$$

Under the same first-order scaling, the input term in the Mamba parameterization has the form

$$\bar{B}_i(x)x^i \simeq \delta(Bx)x^i\sigma(\widetilde{\alpha}_i \cdot x + \widetilde{\beta}_i).$$

Thus

$$\begin{aligned}z_{i,\ell} - z_{i,\ell-1} = \delta A_i z_{i,\ell-1}\sigma(\widetilde{\alpha}_i \cdot x_\ell + \widetilde{\beta}_i)\\ + \delta(Bx_\ell)x_\ell^i\sigma(\widetilde{\alpha}_i \cdot x_\ell + \widetilde{\beta}_i) + o(\delta).\end{aligned}$$

Taking the Euler limit gives

$$dZ_{i,t}^X = A_i Z_{i,t}^X\sigma(\widetilde{\alpha}_i \cdot X_t + \widetilde{\beta}_i)\,dt + (BX_t)X_t^i\sigma(\widetilde{\alpha}_i \cdot X_t + \widetilde{\beta}_i)\,dt.$$

This is the Linear CDE driven by

$$d\omega_{i,t}^X = \sigma(\widetilde{\alpha}_i \cdot X_t + \widetilde{\beta}_i)\,dt, \qquad d\xi_{i,t}^X = X_t X_t^i\sigma(\widetilde{\alpha}_i \cdot X_t + \widetilde{\beta}_i)\,dt,$$

which integrates to the gates stated in the proposition. □

A.2.8 Proposition 4.13: Signature expansion

This is the expansion used in [4, Appendix, expansion for Z_t^X].

Proof. Fix X and let $W_{s,t}^X \in \mathbb{R}^{N\times N}$ be the fundamental solution of the homogeneous equation

$$dW_{s,t}^X = \sum_{i=1}^{d_\omega} A_i W_{s,t}^X\,d\omega_t^{X,i}, \qquad W_{s,s}^X = I_N.$$

Equivalently,

$$W_{s,t}^X = I_N + \sum_{i=1}^{d_\omega}\int_s^t A_i W_{s,r}^X\,d\omega_r^{X,i}.$$

Picard iteration gives

$$W_{s,t}^X = I_N + \sum_{m\geq 1}\sum_{i_1,\dots,i_m=1}^{d_\omega} A_{i_m}\cdots A_{i_1}\int_{s<u_1<\cdots<u_m<t} d\omega_{u_1}^{X,i_1}\cdots d\omega_{u_m}^{X,i_m}.$$

For the word $I = (i_1, \dots, i_m)$, the iterated integral is $\mathrm{Sig}(\omega^X)^I_{s,t}$ and $A_I = A_{i_m} \cdots A_{i_1}$, so

$$W^X_{s,t} = \sum_{I \in \mathbb{W}_{d_\omega}} A_I \, \mathrm{Sig}(\omega^X)^I_{s,t}.$$

Since ω^X has bounded variation, this series is the usual signature series of the linear controlled equation and converges absolutely on compact variation-bounded sets.

Variation of constants for

$$dZ^X_t = \sum_i A_i Z^X_t \, d\omega^{X,i}_t + B \, d\xi^X_t$$

gives

$$Z^X_t = W^X_{0,t} Z_0 + \int_0^t W^X_{s,t} B \, d\xi^X_s.$$

Writing B_j for the jth column of B,

$$B \, d\xi^X_s = \sum_{j=1}^{d_\xi} B_j \, d\xi^{X,j}_s.$$

Substituting the expansion for $W^X_{0,t}$ and $W^X_{s,t}$ gives

$$\begin{aligned} Z^X_t = &\sum_{I \in \mathbb{W}_{d_\omega}} A_I Z_0 \, \mathrm{Sig}(\omega^X)^I_{0,t} \\ &+ \sum_{j=1}^{d_\xi} \sum_{I \in \mathbb{W}_{d_\omega}} A_I B_j \int_0^t \mathrm{Sig}(\omega^X)^I_{s,t} \, d\xi^{X,j}_s. \end{aligned}$$

□

A.2.9 Definition 4.14, Proposition 4.15, and Lemma 4.16: Feature map and RKHS

These are the feature-map and RKHS statements in [4, Appendix, Main Result].

Proof. The full source appendix allows the initial value to depend on the input through a continuous gate $(\cdot)_0 : \mathcal{X} \to \mathbb{R}^{d_0}$. Write the initial condition as $Z^X_0 = C_0 X_0$, and let $C_{0,i}$ be the ith column of C_0. Variation of constants and the Wronskian expansion give

$$\begin{aligned} Z^X_t = &\sum_{i=1}^{d_0} \sum_{I \in \mathbb{W}_{d_\omega}} A_I C_{0,i} \, X^i_0 \, \mathrm{Sig}(\omega^X)^I_{0,t} \\ &+ \sum_{j=1}^{d_\xi} \sum_{I \in \mathbb{W}_{d_\omega}} A_I B_j \int_0^t \mathrm{Sig}(\omega^X)^I_{s,t} \, d\xi^{X,j}_s. \end{aligned}$$

Thus every coordinate of Z^X_t is a linear functional of the feature vector

$$T(X)_{0,t} = \left(X^i_0 \, \mathrm{Sig}(\omega^X)^I_{0,t}, \int_0^t \mathrm{Sig}(\omega^X)^I_{s,t} \, d\xi^{X,j}_s \right)_{i,I,j}.$$

The same object can be defined intrinsically as a tensor-valued CDE. Let $\mathbf{e}_i$ encode the initial-value coordinates, ε^ξ_j the injection coordinates, and ε^ω_k the ω-letters. Then

$$\begin{aligned} T(X)_{0,t} = &\sum_{i=1}^{d_0} X^i_0 \mathbf{e}_i + \sum_{j=1}^{d_\xi} \xi^{X,j}_t \varepsilon^\xi_j \\ &+ \sum_{k=1}^{d_\omega} \int_0^t T(X)_{0,s} \otimes \varepsilon^\omega_k \, d\omega^{X,k}_s. \end{aligned}$$

Iterating this equation gives exactly the displayed coordinates: starting from $\mathbf{e}_i$ produces $X_0^i \operatorname{Sig}(\omega^X)_{0,t}^I$, while starting from ε_j^ξ at time s and then appending ω-letters produces $\int_0^t \operatorname{Sig}(\omega^X)_{s,t}^I \, d\xi_s^{X,j}$.

The RKHS generated by $X \mapsto T(X)_{0,t}$ consists precisely of Hilbert linear functionals on $T(X)_{0,t}$. Splitting such a Hilbert coefficient into the initial-value coordinates and the injection coordinates gives

$$F(X,t) = \sum_i X_0^i \langle \alpha_i, \operatorname{Sig}(\omega^X)_{0,t} \rangle_{\ell^2} + \sum_j \int_0^t \langle \beta_j, \operatorname{Sig}(\omega^X)_{s,t} \rangle_{\ell^2} \, d\xi_s^{X,j}.$$

Conversely, any coefficient families $\alpha_i, \beta_j \in \ell^2(\mathbb{W}_{d_\omega})$ define a Hilbert linear functional on $T(X)_{0,t}$. The norm formula is the standard quotient norm for a feature-map RKHS: take the minimal squared ℓ^2 norm of all Hilbert coefficients representing the same function.

For the kernel, compute the Hilbert inner product of the two feature vectors. The initial-value coordinates give

$$\langle X_0, Y_0 \rangle \langle \operatorname{Sig}(\omega^X)_{0,s}, \operatorname{Sig}(\omega^Y)_{0,t} \rangle_{\ell^2}.$$

The injection coordinates give

$$\int_{\eta=0}^s \int_{\tau=0}^t \langle \operatorname{Sig}(\omega^X)_{\eta,s}, \operatorname{Sig}(\omega^Y)_{\tau,t} \rangle_{\ell^2} \langle d\xi_\eta^X, d\xi_\tau^Y \rangle.$$

This is the signature expression in Lemma 4.16. Alternatively, differentiating the tensor CDE for $T(X)$ against the tensor CDE for $T(Y)$ gives

$$\mathcal{K}^{X,Y}(s,t) = \langle X_0, Y_0 \rangle + \langle \xi_s^X, \xi_t^Y \rangle + \int_0^s \int_0^t \mathcal{K}^{X,Y}(\eta,\tau) \langle d\omega_\eta^X, d\omega_\tau^Y \rangle.$$

□

A.2.10 Theorem 4.17: Dense Linear CDE closure

This is the content of [4, Theorem 4.1 and Appendix, Main Result].

Proof. Let

$$K = \{\omega_{[0,t]}^X : (X,t) \in \mathcal{X} \times [0,1]\}$$

and

$$K_2 = \{\omega_{[s,t]}^X : (X,s,t) \in \mathcal{X} \times [0,1]^2, \ s \le t\}.$$

These are compact because $\mathcal{X}$ is compact, the gates are continuous, and path restriction is continuous in bounded variation topology.

The signature expansion of Proposition 4.13 shows that every linear readout CZ_t^X is a linear functional of the feature coordinates

$$T(X)_{0,t} = \left(\operatorname{Sig}(\omega^X)_{0,t}^I, \int_0^t \operatorname{Sig}(\omega^X)_{s,t}^I \, d\xi_s^{X,j} \right)_{I,j}.$$

Hence the closure of Linear CDE readouts is contained in the uniform closure of linear functionals of T.

We identify that closure. On compact subsets of bounded-variation paths, signature coordinates form an algebra under the shuffle identity:

$$\operatorname{Sig}(\gamma)^I \operatorname{Sig}(\gamma)^J = \sum_{K \in I \sqcup\!\sqcup J} \operatorname{Sig}(\gamma)^K.$$

The added coordinates $\omega_t^{X,1} = t$ and $\omega_t^{X,2} = t^2$ separate the relevant restricted paths uniformly in both X and t; equivalently, they rule out the path identifications that would otherwise require quotienting by tree-like equivalence. Therefore Stone-Weierstrass, applied to the signature algebra, gives: for every continuous

$$\Psi : C_{1,0}([0,1]; \mathbb{R}^{d_\omega}) \to \mathbb{R}$$

and every $\eta > 0$, there is a finite linear combination $P(\mathrm{Sig}(\omega^X)_{0,t})$ such that

$$\sup_{(X,t)\in\mathcal{X}\times[0,1]} |\Psi(\omega^X_{[0,t]}) - P(\mathrm{Sig}(\omega^X)_{0,t})| < \eta.$$

Likewise, for each coordinate of the continuous $\Phi : C_{1,0}([0,1];\mathbb{R}^{d_\omega}) \to \mathbb{R}^{d_\xi}$, there is a finite signature polynomial Q_j satisfying

$$\sup_{\gamma\in K_2} |\Phi_j(\gamma) - Q_j(\mathrm{Sig}(\gamma))| < \eta.$$

Since ξ^X has uniformly bounded variation on the compact set, say $\|\xi^X\|_{1;[0,1]} \le M$, the integral error satisfies

$$\sup_{(X,t)} \left| \int_0^t \left(\Phi(\omega^X_{[s,t]}) - Q(\mathrm{Sig}(\omega^X)_{s,t})\right) \cdot d\xi^X_s \right| \le \eta M.$$

Thus every target of the displayed form is uniformly approximated by a finite linear functional of the coordinates of $T(X)_{0,t}$.

It remains to realize finite linear functionals of T by a finite-dimensional Linear CDE with dense matrices. Choose a finite word set $\mathcal{I}$ large enough to contain the signature coordinates appearing in P and Q. Dense matrices A_i can be chosen as the left-shift representation on the finite word span:

$$A_i e_I = \begin{cases} e_{Ii}, & Ii \in \mathcal{I}, \\ 0, & Ii \notin \mathcal{I}, \end{cases}$$

after identifying the finite word span with $\mathbb{R}^N$. Choosing Z_0 and B to inject the empty-word and ξ coordinates gives a state whose coordinates are precisely the finite coordinates of T, up to the chosen truncation. A final linear functional C forms the required finite combination. Taking η and the truncation error small enough gives the stated ϵ.

The reverse inclusion follows from the first paragraph: every CZ^X_t is already a linear functional of the same signature/integrated-signature feature map, hence belongs to the corresponding closed class. □

A.2.11 Theorem 4.18: Full source closure with initial-value gate

This is [4, Appendix, Theorem Main Result].

Proof. We use the feature map of Definition 4.14. The expansion proved in A.2.9 shows first that every fixed readout $\langle v, Z^X_t\rangle$ is a linear functional of $T(X)_{0,t}$, up to a uniformly vanishing signature-tail truncation. Indeed, for fixed matrices there is $\lambda \ge 0$ such that

$$|v^\top A_I C_{0,i}| + |v^\top A_I B_j| \le \lambda^{|I|}$$

for all words I. Since

$$|\,\mathrm{Sig}(\omega^X)^I_{s,t}| \le \frac{\|\omega^X\|^{|I|}_{1\text{-var};[s,t]}}{|I|!},$$

compactness of $\mathcal{K}$ and continuity of the gates give a uniform bound on the tail

$$\sum_{|I|\ge M} \lambda^{|I|} |\,\mathrm{Sig}(\omega^X)^I_{s,t}|$$

which tends to zero as $M \to \infty$. Hence readouts lie in the uniform closure of the RKHS generated by T.

Conversely, let

$$F(X,t) = \Psi(\omega^X_{[0,t]}) \cdot X_0 + \int_0^t \Phi(\omega^X_{[s,t]}) \cdot d\xi^X_s.$$

The maps

$$(X,t) \mapsto \omega^X_{[0,t]}, \qquad (X,s,t) \mapsto \omega^X_{[s,t]}$$

have compact images. The technical coordinates $\omega_t^{X,1} = t$ and $\omega_t^{X,2} = t^2$ ensure that signatures separate these restricted paths on the relevant compact images, avoiding tree-like identifications in the uniform approximation argument. By the shuffle identity, linear functionals of signatures form an algebra; by Stone-Weierstrass they uniformly approximate the continuous functions Ψ and Φ on those compact images. Uniform bounded variation of ξ^X controls the integral error. Therefore every displayed target F lies in the uniform closure of the RKHS generated by T.

It remains to realize a finite RKHS truncation as a finite Linear CDE. Choose M large enough that replacing each coefficient α_i, β_j by its word-length truncation $\pi_M\alpha_i, \pi_M\beta_j$ changes the target by at most $\epsilon/2$. Let $\mu(M, d)$ be the dimension of the truncated tensor space $T^M(\mathbb{R}^{d_0+d_\xi+d_\omega})$. On this space, let Λ_k denote right tensoring by the kth ω-letter, truncated at level M. The CDE

$$\widetilde{Z}_t^X = \sum_{i=1}^{d_0} X_0^i \mathbf{e}_i + \sum_{j=1}^{d_\xi} \xi_t^{X,j} \varepsilon_j^\xi + \sum_{k=1}^{d_\omega} \int_0^t \Lambda_k \widetilde{Z}_s^X \, d\omega_s^{X,k}$$

is a finite-dimensional Linear CDE of the stated form and its coordinates are exactly the truncated feature vector $\pi_M T(X)_{0,t}$. A readout vector v whose coordinates are the truncated coefficient families $\pi_M\alpha_i, \pi_M\beta_j$ therefore realizes the truncation. This gives the deterministic dense statement.

For the randomized statement, the source replaces the exact truncated-tensor orthogonality above by approximate orthogonality of the random vectors $A_I C_{0,i}$ and $A_I B_j$. Under the LeCun/Gaussian initialization, for fixed truncation level M the relevant inner products satisfy L^2 errors of order $N^{-1/2}$, with constants depending on M and the word lengths. Taking

$$v^{(N)} = \frac{1}{N}\left(\sum_{i,I} \alpha_i^I A_I C_{0,i} + \sum_{j,I} \beta_j^I A_I B_j\right)$$

over the finite set $|I| \le M$ makes $\langle v^{(N)}, Z_t^X\rangle$ converge in L^2, uniformly on the compact set, to the same truncated RKHS functional. Markov's inequality then gives probability tending to one for the uniform approximation event. Combining with the truncation error proves the probability limit.

For the diagonal statement, write $A_i = \mathrm{diag}(v_i)$ and $V = [v_1|\cdots|v_{d_\omega}]$. The homogeneous solution is explicit:

$$W_{s,t}^X = \exp\big(\mathrm{diag}(V(\omega_t^X - \omega_s^X))\big).$$

Thus a readout has the form

$$\psi_N(\omega_t^X) \cdot X_0 + \int_0^t \phi_N(\omega_t^X - \omega_s^X) \cdot d\xi_s^X$$

where ψ_N, ϕ_N are finite sums of exponentials $z \mapsto e^{\langle v,z\rangle}$. These exponential spans form point-separating algebras on compact subsets of $\mathbb{R}^{d_\omega}$, so Stone-Weierstrass gives the displayed diagonal closure. This argument uses only the values and increments of ω^X, so the extra coordinates t, t^2 are unnecessary in the diagonal case. The same expansion also gives the reverse inclusion for fixed diagonal parameters. □

A.2.12 Theorem 4.19: Randomized Linear CDE closure

This is [4, Theorem 4.2 and Appendix, Main Result].

Proof. Fix a target F of the form in Theorem 4.19 and $\epsilon > 0$. By the proof of Theorem 4.17, there is a finite-dimensional feature vector

$$\Theta(X,t) = \left(\mathrm{Sig}(\omega^X)_{0,t}^I, \int_0^t \mathrm{Sig}(\omega^X)_{s,t}^I \, d\xi_s^{X,j}\right)_{(I,j)\in\mathcal{J}}$$

and a coefficient vector a such that

$$\sup_{(X,t)} |F(X,t) - \langle a, \Theta(X,t)\rangle| < \epsilon/2.$$

Thus it suffices to approximate the finite feature span generated by Θ.

For fixed N, the random Linear CDE state $Z_t^X \in \mathbb{R}^N$ gives N random linear functionals of the infinite feature map $T(X)_{0,t}$. The entries of A_i, B, and Z_0 are sampled from nondegenerate Gaussian laws, so for any finite collection of

feature coordinates the induced random coefficient vectors have a distribution with full support on the corresponding finite coefficient space. In particular, for the finite-dimensional subspace containing Θ, the probability that the span of the N readout coordinates fails to approximate the required vector goes to zero as $N \to \infty$. Equivalently,

$$\lim_{N\to\infty} \mathbb{P}\left[\inf_{C\in\mathbb{R}^{1\times N}} \sup_{(X,t)} |\langle a, \Theta(X,t)\rangle - CZ_t^X| \le \epsilon/2\right] = 1.$$

Combining this with the deterministic approximation of F by $\langle a, \Theta\rangle$ gives

$$\lim_{N\to\infty} \mathbb{P}\left[\exists C : \sup_{(X,t)} |F(X,t) - CZ_t^X| \le \epsilon\right] = 1.$$

Only C is chosen after the random draw; the recurrent parameters remain fixed. □

A.2.13 Theorem 4.20: Diagonal Linear CDE closure

This is [4, Theorem 4.3 and Appendix, Diagonal Case].

Proof. Write $A_i = \mathrm{diag}(v_i)$ with $v_i \in \mathbb{R}^N$, and put $V = [v_1|\cdots|v_{d_\omega}] \in \mathbb{R}^{N\times d_\omega}$. Since diagonal matrices commute, the homogeneous Wronskian is explicit:

$$W_{s,t}^X = \exp\big(\mathrm{diag}(V(\omega_t^X - \omega_s^X))\big).$$

Thus variation of constants gives

$$Z_t^X = \exp(\mathrm{diag}(V\omega_t^X))Z_0 + \int_0^t \exp(\mathrm{diag}(V(\omega_t^X - \omega_s^X)))B\, d\xi_s^X.$$

After applying a linear readout C, the first term is a finite linear combination of functions

$$z \mapsto e^{\langle v, z\rangle}, \qquad z = \omega_t^X,$$

and the integral term is a finite linear combination of

$$z \mapsto e^{\langle v, z\rangle}, \qquad z = \omega_t^X - \omega_s^X,$$

integrated against $d\xi_s^X$. Therefore every diagonal readout has the form

$$\psi_N(\omega_t^X) + \int_0^t \phi_N(\omega_t^X - \omega_s^X)\cdot d\xi_s^X$$

with ψ_N, ϕ_N in finite exponential spans. This gives one inclusion in the closure statement.

Conversely, let ψ and ϕ be continuous on the compact sets reached by ω_t^X and $\omega_t^X - \omega_s^X$. The exponential span

$$\mathrm{span}\{z \mapsto e^{\langle v, z\rangle} : v \in \mathbb{R}^{d_\omega}\}$$

is an algebra, contains constants, and separates points; by Stone-Weierstrass it is uniformly dense in continuous functions on those compact sets. Hence there are finite exponential sums ψ_N, ϕ_N with

$$\|\psi - \psi_N\|_\infty < \eta, \qquad \|\phi - \phi_N\|_\infty < \eta.$$

Uniform bounded variation of ξ^X on compact $\mathcal{X}$ gives

$$\sup_{(X,t)} \left|\int_0^t (\phi - \phi_N)(\omega_t^X - \omega_s^X)\cdot d\xi_s^X\right| \le \eta \sup_X \|\xi^X\|_{1;[0,1]}.$$

Choosing η small gives the desired approximation. Each exponential term is implemented by one diagonal coordinate with diagonal vector v, so the finite exponential sums are diagonal Linear CDE readouts.

The same conclusion can be read from signatures: when the A_i commute, $A_I = A_{\sigma(I)}$ for every permutation σ, so the readout can only see the symmetrized signature, which is a function of the increment $\omega_t^X - \omega_s^X$ rather than the ordered path section $\omega_{[s,t]}^X$. □

A.2.14 Corollary 4.21: Mamba channel closure

This is [4, Corollary 4.4 and Appendix, Mamba Case].

Proof. In the Mamba specialization, the model runs independent diagonal systems for the separate channels. The ith channel has one-dimensional gate $\omega^{X,i}$ and input driver $\xi^{X,i}$. By Theorem 4.20, the closure of readouts from the ith subsystem is

$$(X,t) \mapsto \psi_i(\omega_t^{X,i}) + \int_0^t \phi_i(\omega_t^{X,i} - \omega_s^{X,i})\, d\xi_s^{X,i}.$$

The full Mamba-style channel stack concatenates these hidden states and applies a linear output mixing. Linear output mixing produces finite sums of the channelwise closure classes. Taking uniform closures gives exactly

$$\sum_{i=1}^{d_\omega} \psi_i(\omega_t^{X,i}) + \sum_{i=1}^{d_\omega} \int_0^t \phi_i(\omega_t^{X,i} - \omega_s^{X,i})\, d\xi_s^{X,i}.$$

□

A.2.15 Proposition 4.22: Input-driven chained diagonal signature recovery

This is the chaining argument in [4, Appendix, Stability and Chaining of Diagonal Systems].

Proof. We specialize the source level-two recovery condition to the case where the same bounded-variation coordinates of X are available as gate and injection drivers. Let $M = \sup_{X\in\mathcal{X}} \|X\|_{1;[0,1]} < \infty$. The base case is a level-two signature coordinate. For bounded-variation paths,

$$\mathrm{Sig}(X)_{0,t}^{ij} = \int_0^t \int_0^s dX_r^i\, dX_s^j$$

and

$$\begin{aligned} X_t^i X_t^j &= \int_0^t \int_0^t dX_r^i\, dX_s^j \\ &= \int_0^t \int_0^s dX_r^i\, dX_s^j + \int_0^t \int_s^t dX_r^i\, dX_s^j. \end{aligned}$$

The second term is

$$\int_0^t (X_t^i - X_s^i)\, dX_s^j.$$

Therefore

$$\mathrm{Sig}(X)_{0,t}^{ij} = X_t^i X_t^j - \int_0^t (X_t^i - X_s^i)\, dX_s^j.$$

This belongs to the diagonal closure class of Theorem 4.20, so it can be uniformly approximated by a diagonal Linear CDE readout.

Assume inductively that every word I of length k can be recovered by a chain of $k-1$ diagonal CDEs: for every $\eta > 0$ there is a readout $W_{k-1} Z_t^{k-1,X}$ with

$$\sup_{(X,t)} |\,\mathrm{Sig}(X)_{0,t}^I - W_{k-1} Z_t^{k-1,X}| \le \eta.$$

Let Ij be a word of length $k+1$. Then

$$\mathrm{Sig}(X)_{0,t}^{Ij} = \int_0^t \mathrm{Sig}(X)_{0,s}^I\, dX_s^j.$$

Drive the next diagonal CDE by the path

$$\begin{bmatrix} W_{k-1} Z^{k-1,X} \\ X \end{bmatrix}.$$

The level-two argument applied to this augmented driver gives a diagonal readout $vZ_t^{k,X}$ approximating

$$\int_0^t W_{k-1} Z_s^{k-1,X}\, dX_s^j$$

uniformly within any prescribed tolerance. The remaining error is bounded by

$$\begin{aligned}\left|\int_0^t \left(\mathrm{Sig}(X)_{0,s}^I - W_{k-1} Z_s^{k-1,X}\right) dX_s^j\right| &\le \eta \,\|X^j\|_{1;[0,1]} \\ &\le \eta M.\end{aligned}$$

Choosing η and the second approximation tolerance small enough gives the claimed uniform ϵ bound. Induction proves the statement for all $|I| \ge 2$. □

A.2.16 Proposition 4.23: ReLU-gated chained recovery

This is the ReLU-gate observation in [4, Appendix, ReLU activation choice].

Proof. The only possible obstruction is that the Mamba-style gate is constrained to be componentwise nonnegative:

$$\omega_t^X = \int_0^t \mathrm{ReLU}(WX_s + b)\, ds.$$

The algebraic observation is that paired ReLU coordinates preserve the integrated input. If, for each input coordinate j, the gate contains the two rows $e_j^\top$ and $-e_j^\top$ with zero bias, then

$$\mathrm{ReLU}(X_s^j) - \mathrm{ReLU}(-X_s^j) = X_s^j,$$

and hence a fixed linear map L satisfies

$$(L\omega^X)_t^j = \int_0^t X_s^j\, ds = \widetilde{\omega}_t^{X,j}.$$

The more general assumption in the proposition records exactly this linear recoverability.

In the level-two recovery argument, replace every occurrence of ω_t^X inside the functions ψ and ϕ by $L\omega_t^X$. Because L is fixed and linear, the maps

$$\omega_t^X \mapsto \widetilde{\omega}_t^{X,i}\xi_t^{X,j}, \qquad \omega_t^X - \omega_s^X \mapsto \widetilde{\omega}_t^{X,i} - \widetilde{\omega}_s^{X,i}$$

are continuous functions of the admissible diagonal-closure variables. Thus the same diagonal closure theorem approximates

$$\widetilde{\omega}_t^{X,i}\xi_t^{X,j} - \int_0^t (\widetilde{\omega}_t^{X,i} - \widetilde{\omega}_s^{X,i})\, d\xi_s^{X,j}.$$

When the injection driver contains the corresponding $d\widetilde{\omega}^{X,j}$ coordinate, this expression is

$$\int_0^t \int_0^s d\widetilde{\omega}_r^{X,i}\, d\widetilde{\omega}_s^{X,j} = \mathrm{Sig}(\widetilde{\omega}^X)_{0,t}^{ij}.$$

For longer words, use the same induction as in Proposition 4.22. If a previous chain approximates $\mathrm{Sig}(\widetilde{\omega}^X)_{0,s}^I$ uniformly, drive the next diagonal CDE by the previous readout and the available $\widetilde{\omega}^X$ coordinates. The bounded-variation estimate

$$\left|\int_0^t \left(\mathrm{Sig}(\widetilde{\omega}^X)_{0,s}^I - W_{k-1} Z_s^{k-1,X}\right) d\widetilde{\omega}_s^{X,j}\right| \le \eta \,\|\widetilde{\omega}^{X,j}\|_{1;[0,1]}$$

controls the propagated error uniformly on the compact input set. Choosing the induction tolerance small enough gives the required ϵ. □

A.2.17 Proposition 4.24: Chained diagonal approximation

This follows [4, Proposition 4.5 and Appendix, Chaining].

Proof. Let F be in the dense Linear CDE closure class and fix $\epsilon > 0$. By the proof of Theorem 4.17, there is a finite set of signature and integrated-signature coordinates and a linear functional L of those coordinates such that

$$\sup_{(X,t)} |F(X,t) - L(T(X)_{0,t})| < \epsilon/2.$$

Each required signature coordinate can be uniformly approximated by a chained diagonal system by Proposition 4.22. For the integrated coordinates

$$\int_0^t \mathrm{Sig}(\omega^X)^I_{s,t}\, d\xi^{X,j}_s,$$

the same chaining construction is applied to the relevant augmented driver and then integrated through the final diagonal system. Since only finitely many coordinates are required, run the corresponding chains in parallel and concatenate their hidden states. A final linear readout C_k implements the finite linear functional L up to error $\epsilon/2$. Combining the two errors gives

$$\sup_{(X,t)} |F(X,t) - C_k Z^{k,X}_t| \le \epsilon$$

for sufficiently large chain depth and hidden dimensions. □

A.2.18 Proposition 4.25: Path-to-path approximation

This is [4, Proposition 5.1 and Appendix, Path-to-Path].

Proof. Let $G : C_{1,0}([0,1];\mathbb{R}^{d_\omega}) \times [0,1] \to \mathbb{R}$ be continuous and causal. The continuous initial map $X \mapsto X_0$ and the constant coordinate $X^1_0 \equiv 1$ are carried by the source initialization $Z^X_0 = C_0 X_0$. Since $\mathcal{K} \times [0,1]$ is compact and the gates are continuous, $G(\omega^X, t)$ is uniformly continuous in (X,t). Choose a finite mesh

$$0 = s_0 < s_1 < \cdots < s_m = 1$$

so fine that, whenever t lies in the neighboring interval of s_r,

$$|G(\omega^X, t) - G(\omega^X, s_r)| < \epsilon \qquad \text{uniformly in } X \in \mathcal{K}.$$

For each fixed s_r, causality makes $G(\omega^X, s_r)$ a continuous functional of the restricted path $\omega^X_{[0,s_r]}$. Signature density, as in the proof of Theorem 4.17, gives a Linear CDE readout $f_r(X,t) = C_r Z^X_t$ satisfying

$$\sup_{X \in \mathcal{K}} |f_r(X, s_r) - G(\omega^X, s_r)| < \epsilon.$$

By continuity in time and compactness, after refining the mesh if necessary,

$$|f_r(X,t) - f_r(X,s_r)| < \epsilon$$

whenever t is in the neighboring interval of s_r.

Run the finitely many CDEs for $f_0, \ldots, f_m$ in parallel, and include the time coordinate through the gate assumption $\omega^{X,1}_t = t$. The hidden state therefore contains approximations of

$$(t, f_0(X,t), \ldots, f_m(X,t)).$$

On the compact range of this vector, a feed-forward network can uniformly approximate the piecewise linear interpolation map

$$(t, z_0, \ldots, z_m) \mapsto \frac{s_{r+1} - t}{s_{r+1} - s_r} z_r + \frac{t - s_r}{s_{r+1} - s_r} z_{r+1}, \qquad t \in [s_r, s_{r+1}].$$

Combining the time-discretization error, the fixed-time signature approximation error, and the MLP interpolation error yields

$$\sup_{(X,t) \in \mathcal{K} \times [0,1]} |F_{\mathrm{mlp}}(Z^X_t) - G(\omega^X, t)| < C\epsilon$$

for a universal finite constant C. Replacing ϵ by ϵ/C gives the stated bound. □

A.2.19 Proposition 4.26: source exponential-trapezoidal discretization

Proof. On the interval $[\tau_{t-1}, \tau_t]$, variation of constants gives

$$h(\tau_t) = \exp\left(\int_{\tau_{t-1}}^{\tau_t} A(s)\,ds\right) h(\tau_{t-1}) + \int_{\tau_{t-1}}^{\tau_t} \exp\left(\int_{\tau}^{\tau_t} A(s)\,ds\right) B(\tau)x(\tau)\,d\tau.$$

Approximating the state-transition integral by the right endpoint $A_t = A(\tau_t)$ gives

$$h_t \approx e^{\Delta_t A_t} h_{t-1} + \int_{\tau_{t-1}}^{\tau_t} e^{(\tau_t-\tau)A_t} B(\tau)x(\tau)\,d\tau.$$

Set

$$g_t(\tau) = e^{(\tau_t-\tau)A_t} B(\tau)x(\tau).$$

The generalized trapezoidal rule with endpoint weight λ_t is

$$\int_{\tau_{t-1}}^{\tau_t} g_t(\tau)\,d\tau \approx \Delta_t\big((1-\lambda_t)g_t(\tau_{t-1}) + \lambda_t g_t(\tau_t)\big).$$

Since

$$g_t(\tau_{t-1}) = e^{\Delta_t A_t} B_{t-1}x_{t-1}, \qquad g_t(\tau_t) = B_t x_t,$$

the displayed recurrence in Proposition 4.26 follows.

For the local error, suppose g_t has three bounded derivatives on the interval. Taylor expansion around τ_{t-1} gives

$$\int_{\tau_{t-1}}^{\tau_t} g_t(\tau)\,d\tau = \Delta_t g_t(\tau_{t-1}) + \frac{\Delta_t^2}{2} g_t'(\tau_{t-1}) + \frac{\Delta_t^3}{6} g_t''(\tau_{t-1}) + O(\Delta_t^4),$$

while

$$\begin{aligned}\Delta_t\big((1-\lambda_t)g_t(\tau_{t-1}) + \lambda_t g_t(\tau_t)\big) &= \Delta_t g_t(\tau_{t-1}) + \lambda_t \Delta_t^2 g_t'(\tau_{t-1}) \\ &\quad + \frac{\lambda_t \Delta_t^3}{2} g_t''(\tau_{t-1}) + O(\Delta_t^4).\end{aligned}$$

Subtracting yields

$$\left(\frac{1}{2} - \lambda_t\right)\Delta_t^2 g_t'(\tau_{t-1}) + \left(\frac{1}{6} - \frac{\lambda_t}{2}\right)\Delta_t^3 g_t''(\tau_{t-1}) + O(\Delta_t^4).$$

If $\lambda_t = \frac{1}{2} + O(\Delta_t)$, the first term is also $O(\Delta_t^3)$, giving the claimed local order. Setting $\lambda_t = 1$ keeps only the right endpoint and gives the exponential-Euler update. □

A.2.20 Proposition 4.27: reformulated Mamba-3 mask factorization

Proof. Let

$$w_0 = \gamma_0 v_0, \qquad w_t = \beta_t v_{t-1} + \gamma_t v_t \quad (t \geq 1).$$

Then $w = D_{\beta,\gamma} v$ by the definition of the bidiagonal matrix $D_{\beta,\gamma}$. The recurrence becomes

$$h_t = \alpha_t h_{t-1} + w_t, \qquad h_0 = w_0.$$

Unrolling gives

$$h_t = \sum_{s=0}^{t} \left(\prod_{r=s+1}^{t} \alpha_r\right) w_s.$$

This is exactly $h = L_\alpha w = L_\alpha D_{\beta,\gamma} v$.

Equivalently, the coefficient of v_s in h_t is γ_t when $s = t$, is zero when $s > t$, and for $s < t$ is

$$\left(\prod_{r=s+2}^{t} \alpha_r\right)(\gamma_s \alpha_{s+1} + \beta_{s+1}),$$

with empty products equal to 1. This is the lower-triangular mask displayed by Lahoti et al. [26]. The factor L_α is the usual scalar SSM one-semiseparable decay mask, while $D_{\beta,\gamma}$ is two-band. □

A.2.21 Proposition 4.28: source complex-to-real SSM equivalence

Proof. It is enough to check one complex coordinate; stacking coordinates gives the block-diagonal statement. Write the coordinate as $z_t = u_t + iv_t$ and the transition as $a_t + i\theta_t$. Exponential-Euler discretization gives

$$z_t = e^{\Delta_t(a_t+i\theta_t)} z_{t-1} + \Delta_t(B_t + i\widehat{B}_t)x_t.$$

Since

$$e^{\Delta_t(a_t+i\theta_t)} = e^{\Delta_t a_t}\big(\cos(\Delta_t\theta_t) + i\sin(\Delta_t\theta_t)\big),$$

the real and imaginary parts satisfy

$$\begin{bmatrix} u_t \\ v_t \end{bmatrix} = e^{\Delta_t a_t} \begin{bmatrix} \cos(\Delta_t\theta_t) & -\sin(\Delta_t\theta_t) \\ \sin(\Delta_t\theta_t) & \cos(\Delta_t\theta_t) \end{bmatrix} \begin{bmatrix} u_{t-1} \\ v_{t-1} \end{bmatrix} + \Delta_t \begin{bmatrix} B_t \\ \widehat{B}_t \end{bmatrix} x_t.$$

For the output,

$$\Re\big((C_t + i\widehat{C}_t)(u_t + iv_t)\big) = C_t u_t - \widehat{C}_t v_t = \begin{bmatrix} C_t \\ -\widehat{C}_t \end{bmatrix}^\top \begin{bmatrix} u_t \\ v_t \end{bmatrix}.$$

Applying this calculation to each coordinate gives the block-diagonal rotation matrix $\mathcal{R}_t$ and the real input/output projections in the proposition. □

A.2.22 Proposition 4.29: source data-dependent RoPE form

Proof. Let

$$Q_t = \mathcal{R}_0^\top \mathcal{R}_1^\top \cdots \mathcal{R}_t^\top, \qquad Q_{-1} = I.$$

The matrices $\mathcal{R}_t$ are orthogonal block rotations, so $Q_t\mathcal{R}_t = Q_{t-1}$. Multiplying

$$h_t = \alpha_t\mathcal{R}_t h_{t-1} + \beta_t\mathcal{R}_t B_{t-1}x_{t-1} + \gamma_t B_t x_t$$

on the left by Q_t gives

$$Q_t h_t = \alpha_t Q_{t-1}h_{t-1} + \beta_t Q_{t-1}B_{t-1}x_{t-1} + \gamma_t Q_t B_t x_t.$$

With $\widetilde{h}_t = Q_t h_t$, this is the scalar-transition recurrence in the proposition. For the output, orthogonality gives

$$C_t^\top h_t = (Q_t C_t)^\top (Q_t h_t) = (Q_t C_t)^\top \widetilde{h}_t.$$

Thus the transition rotations can be moved into cumulative rotations of the input and output projections. This is the data-dependent RoPE form used by Lahoti et al. [26]. □

A.2.23 Proposition 4.30: reformulated MIMO as SISO components

Proof. The scalar input factor has been absorbed into B_t, matching the source display after writing $\Delta_t B_t$ as B_t. For each input rank j, define the SISO state

$$h_t^{(j)} = \alpha_t h_{t-1}^{(j)} + \gamma_t B_t^{(j)} x_t^{(j)}.$$

Linearity of the recurrence gives

$$h_t = \sum_{j=1}^{\rho} h_t^{(j)} = \alpha_t \sum_{j=1}^{\rho} h_{t-1}^{(j)} + \gamma_t \sum_{j=1}^{\rho} B_t^{(j)} x_t^{(j)}.$$

Applying the ith output projection gives

$$y_t^{(i)} = (C_t^{(i)})^\top h_t = \sum_{j=1}^{\rho} (C_t^{(i)})^\top h_t^{(j)},$$

which is the sum of the ρ SISO maps sharing the transition sequence α. Repeating this for every output rank i yields ρ^2 SISO input-output components.

For the packed matrix form, collect columns into $B_t = [B_t^{(1)}, \ldots, B_t^{(\rho)}]$ and inputs into X_t. The outer product $B_t X_t^\top$ is the sum of the rank-one state-input terms above, and $Y_t = H_t^\top C_t$ collects all output projections. Hence the matrix recurrence is the same algebraic map with the components packed for the MIMO implementation. □

A.2.24 Proposition 4.31: derived affine scan compatibility

Proof. For the real scalar-transition recurrence, define

$$b_t = \beta_t B_{t-1} x_{t-1} + \gamma_t B_t x_t.$$

Then Proposition 4.26 is exactly

$$h_t = \alpha_t h_{t-1} + b_t.$$

This is the affine map

$$F_t(h) = \alpha_t h + b_t.$$

Composition of two affine maps satisfies

$$F_t \circ F_s(h) = \alpha_t \alpha_s h + \alpha_t b_s + b_t,$$

which is the associative product used in Proposition 4.3.

For the realified complex recurrence, put $T_t = \alpha_t \mathcal{R}_t$ and

$$b_t = \beta_t \mathcal{R}_t B_{t-1} x_{t-1} + \gamma_t B_t x_t.$$

The update has the same affine form

$$h_t = T_t h_{t-1} + b_t,$$

now with a matrix linear part. Affine composition is still associative:

$$(T_2, b_2) \circ (T_1, b_1) = (T_2 T_1, T_2 b_1 + b_2).$$

Thus the scan theorem applies once the term b_t, including its dependence on the previous state-input, has been formed. □

A.2.25 Proposition 4.32: derived off-diagonal semiseparable form

Proof. The recurrence is

$$h_t = \alpha_t h_{t-1} + \beta_t v_{t-1} + \gamma_t v_t.$$

For the diagonal coefficient, v_s enters h_s with coefficient γ_s, so $K_{s,s} = \gamma_s$. If $s > t$, causality gives $K_{t,s} = 0$.

For $s < t$, the token v_s first enters h_s with coefficient γ_s and then propagates to h_{s+1} through α_{s+1}; it also enters h_{s+1} directly through the trapezoidal previous-input term with coefficient β_{s+1}. Therefore its total coefficient in h_{s+1} is

$$\alpha_{s+1}\gamma_s + \beta_{s+1}.$$

From time $s+2$ to time t, it is only propagated by the scalar recurrence factors $\alpha_{s+2}, \ldots, \alpha_t$. Hence

$$K_{t,s} = \left(\prod_{r=s+2}^{t} \alpha_r \right) \left(\alpha_{s+1}\gamma_s + \beta_{s+1} \right),$$

with the empty product equal to 1 when $t = s+1$.

Thus each off-diagonal column evolves by the same scalar recurrence law below its first nonzero entry. This is the scalar semiseparable generator form: the generator injected after column s is $\alpha_{s+1}\gamma_s + \beta_{s+1}$, and the propagation factors are α_r. □

A.2.26 Proposition 4.33: derived rotation spectrum and parity sign

Proof. The real rotation matrix

$$R(\vartheta) = \begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix}$$

has complex eigenvectors $(1, -i)^\top$ and $(1, i)^\top$ with eigenvalues $e^{i\vartheta}$ and $e^{-i\vartheta}$. Therefore

$$T_t = e^{\Delta_t a_t} R(\Delta_t \theta_t)$$

has eigenvalues $e^{\Delta_t a_t \pm i\Delta_t \theta_t}$. These eigenvalues are positive real only when the phase is 0 modulo 2π.

For the parity construction, take $a_t = 0$, $\Delta_t = 1$, $\theta_t = \pi x_t$, and remove the input injection. The update is

$$h_t = R(\pi x_t) h_{t-1}.$$

If $x_t = 0$, then $R(\pi x_t) = I$; if $x_t = 1$, then $R(\pi x_t) = R(\pi) = -I$. Thus

$$h_T = \prod_{t=1}^{T} R(\pi x_t) h_0 = (-1)^{\sum_{t=1}^{T} x_t} h_0.$$

With $h_0 = (1, 0)^\top$, this gives

$$h_T = ((-1)^{\sum_{t=1}^{T} x_t}, 0)^\top.$$

The first coordinate is $+1$ for even parity and -1 for odd parity, so a fixed affine threshold maps it to the parity bit. □

A.2.27 Proposition 4.34: derived finite abelian state tracking

Proof. Let $\widehat{G}$ be the character group of G. Since G is finite abelian, the characters form an orthonormal basis for functions $G \to \mathbb{C}$ with respect to the inner product

$$\langle f_1, f_2 \rangle = \frac{1}{|G|} \sum_{g \in G} f_1(g) \overline{f_2(g)}.$$

For each character $\chi \in \widehat{G}$, allocate one complex state coordinate z_t^χ and initialize $z_0^\chi = 1$. For input symbol x_t, define the phase θ_t^χ by

$$e^{i\theta_t^\chi} = \chi(\varphi(x_t)).$$

Use the realified complex Mamba-3 transition with $a_t = 0$, $\Delta_t = 1$, and no input injection. In complex notation this is simply

$$z_t^\chi = e^{i\theta_t^\chi} z_{t-1}^\chi = \chi(\varphi(x_t)) z_{t-1}^\chi.$$

Induction gives

$$z_T^\chi = \prod_{t=1}^{T} \chi(\varphi(x_t)) = \chi\left(\sum_{t=1}^{T} \varphi(x_t) \right) = \chi(g_T),$$

where the second equality uses multiplicativity of characters.

By finite Fourier inversion,

$$\mathbf{1}_A(g) = \sum_{\chi \in \widehat{G}} c_\chi \chi(g), \qquad c_\chi = \frac{1}{|G|} \sum_{a \in A} \overline{\chi(a)}.$$

Therefore the complex linear readout

$$\sum_{\chi \in \widehat{G}} c_\chi z_T^\chi$$

equals $\mathbf{1}_A(g_T)$. Writing each complex coordinate as two real coordinates converts this readout into a real linear readout on $2|\widehat{G}|$ real coordinates. Since $|\widehat{G}| = |G|$, the state dimension is $2|G|$. The output is exactly 0 or 1, so thresholding at $1/2$ recognizes L_A. □

A.2.28 Proposition 4.35: derived componentwise lifting to MIMO

Proof. The identity

$$y^{(i)} = \sum_{j=1}^{\rho} \mathcal{T}_{ij} x^{(j)}$$

is exactly the last display in Proposition 4.30. Because the outer sum is linear, any exact identity $\mathcal{T}_{ij} = \mathcal{S}_{ij}$ for each SISO component gives

$$y^{(i)} = \sum_{j} \mathcal{S}_{ij} x^{(j)}$$

for the MIMO operator.

For the norm statement, use the component bound and triangle inequality:

$$\|y^{(i)}\|_{\mathcal{V}} \le \sum_j \|\mathcal{T}_{ij}x^{(j)}\|_{\mathcal{V}} \le \sum_j L_{ij}\|x^{(j)}\|_{\mathcal{U}}.$$

Let $a_i = \|y^{(i)}\|_{\mathcal{V}}$ and $b_j = \|x^{(j)}\|_{\mathcal{U}}$. Then $a \le Lb$ componentwise with nonnegative vectors, so

$$\|a\|_2 \le \|Lb\|_2 \le \|L\|_{2\to 2}\|b\|_2,$$

which is the desired inequality. □

A.2.29 Proposition 4.36: derived exponential-Euler subfamily containment

Proof. Take any recurrence in $\mathcal{M}_{\mathrm{EE}}^{\mathrm{sc}}$:

$$h_t = e^{\Delta_t A_t}h_{t-1} + \Delta_t B_t x_t, \qquad y_t = C_t^\top h_t.$$

The Mamba-3 exponential-trapezoidal recurrence is

$$h_t = e^{\Delta_t A_t}h_{t-1} + (1-\lambda_t)\Delta_t e^{\Delta_t A_t}B_{t-1}x_{t-1} + \lambda_t\Delta_t B_t x_t.$$

Set $\lambda_t = 1$. Then

$$(1-\lambda_t)\Delta_t e^{\Delta_t A_t}B_{t-1}x_{t-1} = 0, \qquad \lambda_t\Delta_t B_t x_t = \Delta_t B_t x_t,$$

so the two recurrences are identical. The zero-phase real subfamily of the complex formulation has transition

$$e^{\Delta_t A_t}R(0) = e^{\Delta_t A_t}I,$$

so the realified complex representation introduces no additional change. Using the same output vectors C_t gives the same sequence map. Hence every map in $\mathcal{M}_{\mathrm{EE}}^{\mathrm{sc}}$ lies in $\mathcal{M}_3$. □

A.2.30 Proposition 4.37: derived existential theorem transfer

Proof. By Proposition 4.36,

$$\mathcal{M}_{\mathrm{EE}}^{\mathrm{sc}} \subseteq \mathcal{M}_3.$$

If there is an $f \in \mathcal{M}_{\mathrm{EE}}^{\mathrm{sc}}$ with $\mathcal{P}(f)$, then the same f is also an element of $\mathcal{M}_3$. Therefore there is an $f \in \mathcal{M}_3$ with $\mathcal{P}(f)$. □

A.2.31 Proposition 4.38: derived universal theorem transfer criterion

Proof. If $\mathcal{M}_3 \subseteq \mathcal{C}$ and a theorem proves $\mathcal{Q}(f)$ for every $f \in \mathcal{C}$, then every $f \in \mathcal{M}_3$ is also in $\mathcal{C}$, so $\mathcal{Q}(f)$ follows for every Mamba-3 map. The same argument with $\mathcal{M}_3'$ in place of $\mathcal{M}_3$ gives the restricted-subclass statement.

For the positive-real-eigenvalue example, a realified complex Mamba-3 mode has transition eigenvalues

$$e^{\Delta_t a_t \pm i\Delta_t\theta_t}.$$

These eigenvalues are positive real exactly when the phase is an integer multiple of 2π, namely $\Delta_t\theta_t \in 2\pi\mathbb{Z}$. Thus a theorem whose hypotheses require positive real eigenvalues applies directly to that zero-phase restriction. Proposition 4.33 chooses nonzero input-dependent phases, so its construction lies outside that positive-real class. □

A.3 Proofs: State-Space Duality and Attention

A.3.1 Hidden attention

Hidden-attention unrolling of Ali et al. [1]. Starting from

$$h_i = \bar{A}_i h_{i-1} + \bar{B}_i x_i, \qquad h_0 = 0,$$

one obtains by induction

$$h_i = \sum_{j=1}^{i} \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j x_j.$$

Multiplying by C_i gives

$$y_i = \sum_{j=1}^{i} C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j x_j.$$

Thus $y = \widetilde{\alpha} x$ with

$$\widetilde{\alpha}_{ij} = C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j, \qquad j \le i.$$

If $\bar{A}_k$ is diagonal, then the product is coordinatewise:

$$C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j = \sum_{m=1}^{N} C_i[m] \bar{B}_j[m] \prod_{k=j+1}^{i} \bar{A}_k[m,m].$$

This is the decomposition into coordinatewise hidden-attention matrices. □

A.3.2 SSM equals SSS

Proof of the SSM/SSS equivalence in Dao and Gu [7]. Unroll the SSM

$$h_j = A_j h_{j-1} + B_j x_j, \qquad y_j = C_j^\top h_j,$$

with zero initial state. For $i \le j$ the contribution of x_i to h_j is

$$A_j A_{j-1} \cdots A_{i+1} B_i x_i,$$

and hence the contribution to y_j is

$$C_j^\top A_j A_{j-1} \cdots A_{i+1} B_i x_i.$$

Therefore the sequence map is multiplication by the lower triangular matrix

$$M_{ji} = C_j^\top A_j A_{j-1} \cdots A_{i+1} B_i, \qquad i \le j.$$

This is exactly the sequentially semiseparable representation $M = \mathrm{SSS}(A, B, C)$.

To see semiseparability directly, fix a block below the diagonal crossing a split s. For rows $j > s$ and columns $i \le s$,

$$M_{ji} = \underbrace{C_j^\top A_j \cdots A_{s+1}}_{U_j} \underbrace{A_s \cdots A_{i+1} B_i}_{V_i}.$$

The block factors as UV through an N-dimensional state space, so its rank is at most N. This proves that an N-SSS representation gives an N-semiseparable matrix. The converse, that every N-semiseparable matrix admits an N-SSS representation, is the structured-matrix representation result used by Dao and Gu [7] and restated constructively by Hu et al. [20]. □

A.3.3 Semiseparable matrix-vector multiplication

Proof of Dao and Gu [7], Proposition 3.6. The proposition is a structured-matrix fact for semiseparable matrices, used as a black-box input by Dao and Gu [7]. In compressed generator form, an N-semiseparable $T \times T$ matrix stores $O(N)$ row and column generators per time index, hence $O(NT)$ parameters overall. Matrix-vector multiplication is then a single causal scan over the sequence: the scan state records the N-dimensional summary of all previous columns that can affect the current row. At step t, the generator update costs $O(N)$ operations and the output contraction also costs $O(N)$. Summing over T steps gives $O(NT)$ time and space.

If one starts from a dense sequentially semiseparable representation with arbitrary $A_t \in \mathbb{R}^{N \times N}$, converting it to the compressed generator representation may require preprocessing. This is why the following efficiency theorem explicitly excludes preprocessing from the stated runtime. □

A.3.4 Efficient SSM computation

Proof of Dao and Gu [7], Theorem 3.7. Let M be the sequence matrix induced by an SSM of state size N on a sequence of length T. By the SSM/SSS equivalence in Appendix A.3.2, M is an N-semiseparable matrix in sequentially semiseparable representation.

By Appendix A.3.3, an N-semiseparable $T \times T$ matrix admits an $O(NT)$-parameter compressed representation and supports matrix-vector multiplication in $O(NT)$ time and space. Applying this multiplication routine to Mx computes the full output sequence of the SSM. The theorem excludes possible preprocessing needed to convert an arbitrary representation into the compressed one; after that representation is available, the sequence computation costs $O(TN)$. □

A.3.5 Scalar-identity SSD duality

Scalar-identity SSM as structured masked attention. Assume $A_t = a_t I_N$. The SSM sequence matrix is

$$M_{ji} = C_j^\top (a_j I)(a_{j-1} I) \cdots (a_{i+1} I) B_i = \left(\prod_{k=i+1}^{j} a_k \right) C_j^\top B_i.$$

Define the 1-semiseparable causal mask

$$L_{ji} = \prod_{k=i+1}^{j} a_k, \qquad i \le j,$$

and $L_{ji} = 0$ for $i > j$. If the rows of C and B are $C_j^\top$ and $B_i^\top$, then $(CB^\top)_{ji} = C_j^\top B_i$. Hence

$$M = L \odot (CB^\top).$$

Multiplication Mx may be computed either by the recurrent SSM scan or by materializing the quadratic structured masked-attention matrix. These are the two dual forms of the same transformation.

The corollary for 1-SS structured masked attention follows by reversing this calculation. The bottleneck multiplication by L is a scalar SSM recurrence; placing the query-key outer product around that mask gives the diagonal SSM with scalar-identity transition. □

A.3.6 Autoregressive structured masked attention

Proof of Dao and Gu [7], Theorem 5.2. Let $y = Lx$ be an autoregressive linear transformation of order k in the sense that

$$y_t = \mu_t x_t + \ell_{t1} y_{t-1} + \cdots + \ell_{tk} y_{t-k}.$$

Moving the previous-output terms to the left gives

$$y_t - \ell_{t1} y_{t-1} - \cdots - \ell_{tk} y_{t-k} = \mu_t x_t.$$

Vectorizing over t gives

$$By = D_\mu x,$$

where B is lower triangular and $(k+1)$-banded, while D_μ is diagonal. Thus $L = B^{-1} D_\mu$.

It remains to show that B^{-1} is semiseparable with rank controlled by k. Fix a split time s, put the input to zero after the split, and look at future outputs $y_{s+1}, y_{s+2}, \ldots$. For $t > s$ the recursion becomes homogeneous:

$$y_t = \ell_{t1} y_{t-1} + \cdots + \ell_{tk} y_{t-k}.$$

Hence the entire future vector $(y_{s+1}, y_{s+2}, \ldots)$ is determined by the boundary state

$$r_s = (y_s, y_{s-1}, \ldots, y_{s-k+1}) \in \mathbb{R}^k.$$

The map from past inputs $x_1, \ldots, x_s$ to this boundary state is linear, and the map from the boundary state to future outputs is also linear. Thus every strictly lower off-diagonal block of L crossing the split factors as

$$x_{1:s} \longmapsto r_s \in \mathbb{R}^k \longmapsto y_{s+1:T}.$$

Its rank is at most k. Since the split was arbitrary, L is k-semiseparable, and by the SSM/SSS equivalence it has a finite-order SSM representation. □

A.3.7 SSD block algorithm

Proof of the SSD algorithm of Dao and Gu [7]. The SSD sequence matrix is semiseparable and also has the scalar-identity attention form

$$M = L \odot (CB^\top).$$

Partition the sequence into contiguous blocks of length Q. For two positions in the same block, the output can be computed directly from the quadratic attention form because the corresponding submatrix has size $Q \times Q$. Summing over all diagonal blocks costs $O((T/Q)Q^2N) = O(TQN)$ when the value/head dimension is N.

For positions in different blocks, the semiseparable factorization lets all communication pass through boundary states. If a split point s separates input position i and output position j, then

$$M_{ji} = \underbrace{C_j^\top A_j \cdots A_{s+1}}_{\text{right factor}} \underbrace{A_s \cdots A_{i+1} B_i}_{\text{left factor}}, \qquad i \le s < j.$$

For a whole block of past inputs, the left factors can be summarized by one state-like N-dimensional quantity at the block boundary. The block summaries are then propagated by the associative scan

$$h_{\text{end}(b)} = A_{\text{block}(b)} h_{\text{end}(b-1)} + u_{\text{block}(b)},$$

where $A_{\text{block}(b)}$ is the product of transitions across block b and $u_{\text{block}(b)}$ is the contribution of inputs inside that block. The right factors inside later blocks contract these boundary states to outputs. Hence the off-diagonal part is computed by block scans and blockwise matrix multiplications rather than by materializing the full $T \times T$ matrix.

Choosing the SSD regime $P = N$ and the source block size $Q = N$, the diagonal attention work, the state propagation between blocks, and the contractions from boundary states all have total training cost $O(TN^2)$ and use matrix-multiplication-dominated primitives. The source assumes $Q \mid T$ for notation; a shorter final block only changes constants. During autoregressive inference one only stores the current state summary for each head, an $N \times N$ object in this parameterization, and updates it once per new token. This gives $O(TN)$ total inference FLOPs and $O(N^2)$ inference memory. □

A.3.8 Mamba S6 head pattern

Head-pattern translation for Dao and Gu [7], Proposition 7.2. In the SSD vocabulary, a sequence transformation has state size N, head dimension P, and number of heads H. The tensor X is the value-like input, while B and C play the key-like expansion and query-like contraction roles in the attention-dual matrix $L \odot (CB^\top)$.

The original Mamba S6 layer applies one scalar SSM channel to each expanded input channel. Thus each channel is a separate sequence head with $P = 1$, and the decay or transition parameter A is independent across those channels. The input-dependent projections B and C are not given separate copies for every scalar channel; they are shared across the input channels of X. In the head-pattern terminology of Dao and Gu [7], this is the multi-input SSM pattern, or equivalently the multi-value-attention analog. □

A.3.9 Diagonal SSD

Diagonal state-space duality of Hu et al. [20]. Let $A_t = \operatorname{diag}(a_t^{(1)}, \dots, a_t^{(N)})$. The SSM sequence matrix satisfies

$$M_{ji} = c_j^\top A_j A_{j-1} \cdots A_{i+1} b_i = \sum_{n=1}^{N} c_j^{(n)} b_i^{(n)} \prod_{k=i+1}^{j} a_k^{(n)}.$$

For each coordinate n, define the 1-SS mask

$$L_{ji}^{(n)} = \prod_{k=i+1}^{j} a_k^{(n)}$$

and the vectors $c^{(n)} = (c_1^{(n)}, \dots, c_T^{(n)})^\top$ and $b^{(n)} = (b_1^{(n)}, \dots, b_T^{(n)})^\top$. Then

$$M = \sum_{n=1}^{N} L^{(n)} \odot \left(c^{(n)} (b^{(n)})^\top\right).$$

This is the diagonal SSM as a sum of N one-semiseparable masked-attention heads.

If every A_t is invertible, define cumulative products

$$r_t^{(n)} = \prod_{k=1}^{t} a_k^{(n)}.$$

Then for $i \le j$,

$$\prod_{k=i+1}^{j} a_k^{(n)} = \frac{r_j^{(n)}}{r_i^{(n)}}.$$

Absorb $r_j^{(n)}$ into the output vector and $(r_i^{(n)})^{-1}$ into the input vector:

$$c_j'^{(n)} = c_j^{(n)} r_j^{(n)}, \qquad b_i'^{(n)} = b_i^{(n)} (r_i^{(n)})^{-1}.$$

Then

$$M_{ji} = \sum_{n=1}^{N} c_j'^{(n)} b_i'^{(n)} = (C' B'^{\top})_{ji}$$

on the causal lower triangle, which is $M = 1\,\mathrm{SS}(\mathbf{1}) \odot (C'B'^{\top})$. □

A.3.10 General duality condition and boundary

Proof of Hu et al. [20], Theorem 4.1 and Section 5. The constructive semiseparable theorem says that a lower triangular matrix is N-semiseparable if and only if it has an N-SSS representation. Therefore one may discuss whether a general SSM has a 1-SS masked-attention dual by looking directly at its N-semiseparable sequence matrix M.

For a lower triangular M, call column t new if the tail $M_{t:T,t}$ is not in the span of the previous tails $M_{t:T,1}, \dots, M_{t:T,t-1}$. This is the number of genuinely new column directions introduced while scanning left to right.

For a fine 1-SS mask $L = 1\,\mathrm{SS}(a)$, the representation

$$M = L \odot (QK^{\top}), \qquad Q, K \in \mathbb{R}^{T \times N},$$

can introduce at most N new columns. Indeed, since L is fine, define $r_t = \prod_{\ell=1}^{t} a_\ell$. Multiplying row t by r_t^{-1} and column t by r_t preserves which columns are new, and transforms the lower triangular part of M into the lower triangular part of $QK^{\top}$. If M had $N+1$ new columns at indices $t_1 < \cdots < t_{N+1}$, then the full columns $(QK^{\top})_{:,t_1}, \dots, (QK^{\top})_{:,t_{N+1}}$ would be linearly independent: taking a nontrivial dependence and looking at the largest active index contradicts newness of the corresponding tail. This is impossible because $\mathrm{rank}(QK^{\top}) \le N$.

Conversely, suppose M has at most N new columns. Fill the entries above the diagonal from left to right. If column t is new, set its entries above the diagonal to zero. If it is not new, choose coefficients $c_1, \dots, c_{t-1}$ such that

$$M_{t:T,t} = \sum_{s<t} c_s M_{t:T,s},$$

and set the above-diagonal part of the filled column to $\sum_{s<t} c_s M'_{1:t-1,s}$. Induction on t shows that the filled matrix M' has rank at most the number of new columns, hence at most N. Therefore $M' = QK^{\top}$ for some $Q, K \in \mathbb{R}^{T \times N}$, and on the lower triangle

$$M = 1\,\mathrm{SS}(\mathbf{1}) \odot (QK^{\top}).$$

This proves the fine-mask equivalence.

If the 1-SS mask has zeros, the zeros split the lower triangular matrix into diagonal blocks: entries crossing a zero transition in the mask vanish. On each block the induced mask is fine, so the argument above applies blockwise. Hence a general 1-SS masked-attention dual exists exactly when all nonzero entries of M lie in diagonal blocks and each block has at most N new columns.

For the softmax obstruction, take $V \in \mathbb{R}^{T \times T}$ with $V_{ij} = ij$. This matrix has rank one. But

$$\mathrm{Softmax}(V)_{ij} = \frac{\exp(ij)}{\sum_{\ell=1}^{T} \exp(i\ell)}$$

is a row-rescaled Vandermonde matrix in the bases $e^1, e^2, \ldots, e^T$, hence has full rank; its square submatrices are full rank for the same Vandermonde reason. A finite-dimensional SSM dual would force fixed semiseparable rank. Generic softmax attention therefore does not fit the finite-state SSD duality without additional structure.

The paper also gives an SSM-side obstruction. For $N \geq 2$, there are SSM parameters whose sequence matrix is $M = I + E_{T1}$. If $M = L \odot (QK^\top)$ with a 1-SS mask, then $M_{T1} \neq 0$ forces the mask entries connecting the first and last position to be nonzero. The zero lower triangular entries before the last row then force many entries of $QK^\top$ to vanish, while the diagonal entries remain nonzero. This implies $\operatorname{rank}(QK^\top) \geq T-1$, contradicting $\operatorname{rank}(QK^\top) \leq N$ when T is large relative to N. □

A.4 Proofs: Expressivity and Formal Languages

A.4.1 Flip-Flop construction

Proof of the Flip-Flop construction of Sarrof et al. [35]. The Flip-Flop language uses instruction symbols r, w, i and data symbols $0, 1$. A valid read after r must reproduce the data bit following the most recent preceding write instruction w.

The first SSM layer records the previous instruction while forwarding the current data token. In the scalar form used by Sarrof et al. [35], one can set

$$A(r) = A(w) = A(i) = 0, \qquad A(0) = A(1) = 1, \qquad B(0) = B(1) = 0,$$

and choose $B(r), B(w), B(i)$ to encode the instruction token. Thus, after a data token, the layer output contains both the current bit and the instruction that immediately preceded it.

The second layer is a set-reset memory for the last written data bit. If the first-layer output indicates that the preceding instruction was w, the gate overwrites the hidden state with the current bit. If the preceding instruction was r or i, the gate preserves the previous hidden state. The readout then uses the current instruction and the stored bit to list the legal next symbols: after a read instruction, only the stored bit is legal; after write or ignore, the legality condition does not force the read bit. All hidden values range over a fixed finite set, so the construction is finite-precision and works for all lengths. □

A.4.2 Star-free characterization

Proof of the star-free characterization of Sarrof et al. [35]. The parity lower bound is the base obstruction. On inputs 1^t, a nonnegative finite-precision SSM layer has no oscillatory transition. For one coordinate of a recurrent state, after the input has stabilized the update has the scalar affine form

$$h_t = \alpha h_{t-1} + \beta, \qquad \alpha \geq 0.$$

Thus

$$h_t = \alpha^t h_0 + \beta \sum_{s=0}^{t-1} \alpha^s.$$

If $\alpha < 1$ the sequence converges; if $\alpha = 1$ it is affine in t; and if $\alpha > 1$ it diverges exponentially unless the corresponding coefficient vanishes. Linear maps preserve this trichotomy. Sigmoid gates in GLU components converge to finite limits on such inputs, and SwiGLU components are products of terms that converge, diverge polynomially, or diverge exponentially. After normalization, the dominating coordinates converge to a fixed direction and the dominated ones converge to zero; finite precision then makes the top-layer representation ultimately constant. PARITY alternates on the same subsequence, so no nonnegative finite-precision SSM of this form recognizes PARITY at arbitrary lengths.

For a non-star-free regular language, the McNaughton–Papert and Schutzenberger theory gives words u, v, w such that membership of $uv^t w$ is nonconstant periodic in t. Repeating the same convergence argument on the v-loop gives, for each position i inside the period and each layer k,

$$h^{(k)}_{t|v|+i} = \alpha^{(k)}_i \odot h^{(k)}_{(t-1)|v|+i} + \beta^{(k)}_i$$

for all sufficiently large t, because the lower-layer states have already become stationary by induction. Hence the states along v^t and then along the fixed suffix w are ultimately constant at finite precision. This contradicts the

required dependence of membership on t mod k. Thus nonnegative finite-precision SSMs cannot recognize non-star-free regular languages.

For the positive direction, every star-free language has an aperiodic syntactic monoid. The Krohn–Rhodes decomposition, together with Schutzenberger's aperiodicity characterization, represents the corresponding automaton as an iterated cascade of set-reset automata. A set-reset automaton has transitions of the form "preserve the current state" or "reset to a state determined by the current symbol." An SSM layer implements this with a nonnegative gate:

$$h_t = A(x_t)h_{t-1} + B(x_t),$$

where $A(x_t) = I$ and $B(x_t) = 0$ for preserve symbols, while $A(x_t) = 0$ and $B(x_t) = e_{\delta(x_t)}$ for reset symbols.

It remains to implement cascades. If automaton $\mathcal{A}_1$ feeds automaton $\mathcal{A}_2$, then $\mathcal{A}_2$ needs the pair consisting of the current input symbol and the previous state of $\mathcal{A}_1$. A one-layer delay gadget stores the previous symbol/state at finite precision by using a contracting update such as $A = 1/4$, which leaves a constant decoding margin. Stacking this delay layer between the SSM for $\mathcal{A}_1$ and the SSM for $\mathcal{A}_2$ implements the cascade product. Iterating the construction implements the full aperiodic cascade. The final readout maps the finite automaton state to the set of legal next symbols, so every star-free regular language is predictively modeled at finite precision and at all input lengths. □

A.4.3 Counter languages and bounded Dyck

Counter and bounded-Dyck constructions of Sarrof et al. [35]. For a counter language, let $c : \Sigma \to \mathbb{Z}^m$ be the vector of counter increments induced by each symbol. The SSM update

$$h_t = h_{t-1} + c(x_t)$$

is obtained by taking $A \equiv 1$ and $B(x_t) = c(x_t)$. A readout that distinguishes negative, zero, and positive counter values gives the legal next symbols for Dyck-1, Shuffle-Dyck-k, $a^n b^n$, $a^n b^n c^n$, and the other listed counter languages. The state values may grow with input length; the construction uses the source's finite-precision convention with unbounded integer bits and bounded fractional precision, and is therefore different from the bounded finite-state constructions.

For $\mathrm{Dyck}_{K,h}$, the first layer tracks the current depth up to h by the same counter mechanism. The second layer keeps, for each depth $1, \ldots, h$, the type of the last open bracket at that depth. This is a set-reset memory: an opening bracket at depth d writes its type into the dth register, and other registers persist. Since each register stores one of K types, it needs $O(\log K)$ finite-precision coordinates; over h depths the width is $O(h \log K)$. The readout allows an opening bracket when the depth is below h, and allows a closing bracket exactly when it matches the stored type at the current depth. □

A.4.4 n-gram SSM construction

Proof of the n-gram SSM construction of Nandakumar et al. [30]. First, if an SSM has transition matrix A with $A^n = 0$, then its recurrent state satisfies

$$h_T = \sum_{j=0}^{T-1} A^j B x_{T-j}.$$

All summands with $j \geq n$ vanish, so h_T depends only on $x_T, x_{T-1}, \ldots, x_{T-n+1}$. This proves the finite-context claim.

Second, let $(x_i, y_i)_{i=1}^K$ be distinct finite input-output pairs. The paper proves that, for a generic choice of A and B subject to the desired nilpotency constraint, the hidden vectors $v_i = h(x_i)$ can be made linearly independent after applying the hidden bias and ReLU. If V is the matrix with rows $v_i^\top$ and Y is the matrix with rows $y_i^\top$, then independence gives a linear output map C satisfying

$$VC = Y.$$

Thus a one-hidden-layer SSM with K hidden units memorizes the K desired associations.

For the n-gram theorem, the finite set P contains the admissible $(n-1)$-token contexts. For each context $p \in P$, choose a strictly positive probability vector s_p within ϵ in ℓ^1 of the n-gram next token distribution, and choose

logits $\bar{s}_p$ with $\mathrm{Softmax}(\bar{s}_p) = s_p$. The nilpotent transition ensures that the SSM state depends only on the last $n-1$ tokens; the memorization result chooses the output map so that this context is sent to $\bar{s}_p$. Applying softmax gives an ϵ-equivalent next-token distribution on L. □

A.4.5 Polynomial selectivity

Polynomial-selectivity calculations. For the simplified scalar S6 model of Cohen-Karlik et al. [5], take the input-dependent transition to be linear in the token. After unrolling, the last-token output has the form

$$y_L = \sum_{j=1}^{L} c_j x_L^2 x_j^2 \prod_{k=j+1}^{L-1} x_k,$$

after absorbing fixed coefficients into c_j. These monomials have degree growing linearly with L. A scalar single-head causal linear-attention layer has last output of the form

$$y_L = \sum_{j=1}^{L} c_j x_j^2 x_L,$$

which has degree 3. Stacking N polynomial layers multiplies degrees, so an N-layer linear-attention stack has degree at most 3^N. To represent the linearly growing degree realized by one S6 channel, one needs $3^N \gtrsim L$, so the required number of stacked layers grows logarithmically with L.

The reverse containment for one attention head is obtained by choosing the S6 parameters so that the recurrence accumulates exactly the same causal quadratic value terms as the linear-attention head. This is a representation statement for the simplified polynomial S6 class.

For the three-layer monomial lemma, work in the simplified source setting with state size $N = 1$ and without SiLU activations. The first block isolates the chosen position j. A positional-encoding channel is set to the indicator of position j, the first-half input channels duplicate the scalar input, and the linear maps around the S6 block are identities. With the S6 parameters chosen as $\bar{A} = 0$ and $\bar{B} = \bar{C} = 1$, the output sequence has x_j only at the selected position and zero elsewhere.

The second block duplicates this selected value. Its gate is set to one on the positions used for the construction and zero afterward, while the S6 recurrence with $\bar{A} = \bar{B} = \bar{C} = 1$ copies the selected scalar into the needed run of positions. If the desired degree P exceeds the available sequence room, the zero-padding component lengthens the sequence before this copying step.

The third block multiplies the duplicates. It runs two identical S6 channels on two sequences that differ by inserting one initial zero before the run of copies, and then subtracts the two channel outputs in the final linear map. The two accumulated product sums cancel telescopically, leaving exactly the last product term x_j^P. The coefficient c is inserted by scaling the corresponding output channel.

For the four-layer Mamba polynomial theorem, write a bounded-degree polynomial as a finite sum of monomials

$$P(x) = \sum_{i=1}^{T} c_i \prod_{j=1}^{L} \alpha_{i,j} x_j^{p_{i,j}}.$$

By the monomial lemma, the first three blocks create the univariate factors $s_j = \alpha_{i,j} x_j^{p_{i,j}}$ in separate channels. The fourth S6 block, with system matrices equal to one, forms the product $\prod_j s_j$ by the same product-accumulation construction. The final linear output map weights these product channels by c_i and sums them, realizing the polynomial in the simplified model. □

A.4.6 FSA mapping and diagonal commutativity

FSA mapping and diagonal commutativity in Terzić et al. [40]. For an automaton $\mathcal{A} = (Q, \Sigma, q_{\mathrm{init}}, \delta)$ with one-hot encodings e_q, define

$$A(\sigma) = \sum_{q \in Q} e_{\delta(q,\sigma)} e_q^\top.$$

If $\mathbf{q}_{t-1} = e_q$, then $A(x_t)\mathbf{q}_{t-1} = e_{\delta(q,x_t)}$. Induction gives $\mathbf{q}_t = e_{\delta(q_{\text{init}},x_1\cdots x_t)}$, so the recurrence emulates the automaton.

Now suppose all $A(x)$ are simultaneously diagonalizable: $A(x) = W\Lambda(x)W^{-1}$ with diagonal $\Lambda(x)$. In the changed basis $\widetilde{\mathbf{q}}_t = W^{-1}\mathbf{q}_t$,

$$\widetilde{\mathbf{q}}_T = \Lambda(x_T)\cdots\Lambda(x_1)\widetilde{\mathbf{q}}_0.$$

Diagonal matrices commute, so this product is invariant under reordering of the input symbols. Any automaton emulated by this no-B single-layer mapping therefore has order-insensitive transition behavior, which is the commutative case isolated in the paper. □

A.4.7 Proposition 6.13: PD-SSM monoid and linear multiplication

This follows Properties 1–2 of Terzić et al. [41].

Proof. Represent a matrix $A \in \mathbb{H}^{N\times N}$ by a column map $p_A : \{1,\dots,N\} \to \{1,\dots,N\}$ and nonzero column weights a_j:

$$Ae_j = a_j e_{p_A(j)}.$$

For $B \in \mathbb{H}^{N\times N}$ with data (p_B, b_j),

$$ABe_j = A(b_j e_{p_B(j)}) = b_j a_{p_B(j)} e_{p_A(p_B(j))}.$$

Thus AB again has exactly one nonzero in each column, with column map $p_A \circ p_B$ and weight $b_j a_{p_B(j)}$. The identity matrix belongs to $\mathbb{H}^{N\times N}$, and associativity is inherited from ordinary matrix multiplication, so the class is a monoid.

The displayed formula also gives the $\Theta(N)$ multiplication algorithm: for each column j, look up $p_B(j)$, compose the two indices, and multiply the two scalar weights. This is constant work per column. □

A.4.8 Proposition 6.14: PD-SSM stability

This is the stability bound stated by Terzić et al. [41].

Proof. Unroll the recurrence:

$$h_t = A_{t:1}h_0 + \sum_{s=1}^{t} A_{t:s+1}b_s, \qquad A_{r:s} = A_r A_{r-1}\cdots A_s.$$

Because each $A_i \in \mathbb{H}^{N\times N}$ and $\|A_i\|_\infty \le 1-\varepsilon$, the product $A_{t:s}$ is again a one-hot-column matrix whose nonzero column weights have magnitude at most $(1-\varepsilon)^{t-s+1}$. Hence

$$\|A_{t:s}\|_2 \le \sqrt{N}(1-\varepsilon)^{t-s+1}.$$

Therefore

$$\|h_t\|_2 \le \sqrt{N}(1-\varepsilon)^t B + \sum_{s=1}^{t}\sqrt{N}(1-\varepsilon)^{t-s}B$$

$$\le \sqrt{N}B\sum_{r=0}^{t}(1-\varepsilon)^r \le \frac{\sqrt{N}B}{\varepsilon}.$$

□

A.4.9 Theorem 6.15: PD-SSM finite-state tracking

This collects Propositions 1–3 of Terzić et al. [41].

Proof. For exact FSA emulation, let $\mathcal{A} = (Q, \Sigma, q_{\text{init}}, \delta, F)$ with $|Q| = N$. Encode state q as $e_q \in \mathbb{R}^N$. For each input symbol σ, define

$$P(\sigma)e_q = e_{\delta(q,\sigma)}.$$

The matrix $P(\sigma)$ has exactly one nonzero in each column, hence belongs to $\mathbb{H}^{N\times N}$. Set $D(\sigma) = I$ and $b_t = 0$. Then the PD-SSM recurrence satisfies

$$h_t = P(x_t)h_{t-1} = e_{\delta(q_{\text{init}},x_1\cdots x_t)}$$

by induction. A linear readout with rows indexed by states recovers the current one-hot state or the accepting indicator.

For the near-optimality statement under unique encodings, choose an N-state automaton whose N states must be distinguished by the readout and whose transition graph can reach all states from the initial state. If a single-layer SSM with state dimension $d < N-1$ used unique state encodings for all reachable automaton states, those N encodings would lie in a d-dimensional affine space. Require the readout to return the one-hot state vector $e_q \in \mathbb{R}^N$. A linear readout with bias is an affine map $R(v) = Wv + b$, so the image of the N encodings has affine dimension at most d. But $\{e_q : q \in Q\}$ is the vertex set of an $(N-1)$-simplex: the differences $e_q - e_{q_0}$ for $q \neq q_0$ are linearly independent, so their affine hull has dimension $N-1$. Therefore $d \geq N-1$.

The scalar arbitrary-precision construction shows why the unique-encoding and ordinary-readout assumptions matter. Let the scalar recurrence be

$$x_{t+1} = x_t + u_t\sqrt{p_t},$$

where p_t is the tth prime and $u_t \in \mathbb{Q}$ encodes the input symbol. The numbers $\{\sqrt{p_t}\}$ are linearly independent over $\mathbb{Q}$, so two different finite input sequences yield different scalar states. A lookup table can therefore map each reachable scalar to the corresponding FSA state. This is an exact representation only because arbitrary precision and a nonstandard lookup readout are allowed. □

A.4.10 DCD group tracking

DCD group-tracking characterization of Shakerinava et al. [36]. The source abstraction assumes the transition map A, additive map b, and decoder are universal function approximators; the characterization is therefore about the diagonal complex recurrence structure rather than about a particular finite neural parameterization.

For the one-layer sufficiency direction, write a finite Abelian group as

$$G \simeq C_{k_1} \times \cdots \times C_{k_m}.$$

Represent $g = (r_1, \ldots, r_m)$ by the diagonal matrix

$$\Lambda(g) = \operatorname{diag}\left(e^{2\pi i r_1/k_1}, \ldots, e^{2\pi i r_m/k_m}\right).$$

With $h_0 = \mathbf{1}$ and $b \equiv 0$, the product of input matrices records the running group product. A finite-precision decoder distinguishes the finitely many roots of unity.

For necessity, unrolling a diagonal affine recurrence over a sequence $\bar{x}$ gives

$$h_T = \lambda(\bar{x}) \odot h_0 + b(\bar{x}), \qquad \lambda(\bar{x}) = \prod_{t=1}^{T} \lambda(x_t).$$

Eliminate coordinates that contract, expand, or translate: such coordinates can be replaced by constants without losing exact finite-precision tracking. The remaining useful coordinates are neutral rotations. If two inputs rotate about distinct centers in a coordinate, then composing one rotation with the conjugate rotation produces a nonzero translation, and repeated identity-product words force divergence. Hence all inputs share the same center in each useful coordinate. Rotations about a common center commute, so the represented group operation is commutative.

For k layers, the same coordinate simplification is applied layer by layer. Finite precision makes the reachable state graph finite. Inside a strongly connected component, group elements that share the same first-layer state form cosets of a normal subgroup $G_{k-1} \trianglelefteq G$, and commutativity of the first-layer rotations gives an Abelian quotient G/G_{k-1}. Restricting to input words that keep the first layer fixed leaves a $(k-1)$-layer DCD SSM that tracks G_{k-1}. Iterating yields

$$G = G_k \trianglerighteq G_{k-1} \trianglerighteq \cdots \trianglerighteq G_0 = \{e\}, \qquad G_{i+1}/G_i \text{ Abelian.}$$

Conversely, given such a chain, construct the first layer to track the Abelian quotient G/G_{k-1}. Choose a section $s : G/G_{k-1} \to G$. Every state can be written as $n\,s(h)$ with $n \in G_{k-1}$. When multiplying by an input $g = s(a)n_g$, the quotient layer updates $h \mapsto ha$ and sends the correction

$$\kappa(h, g) = s(ha)^{-1}s(h)s(a)n_g s(ha) \in G_{k-1}$$

to the remaining layers, using normality of G_{k-1}. By induction the remaining $k-1$ layers track this correction inside G_{k-1}. □

A.4.11 Eigenvalue obstructions and Householder constructions

Eigenvalue obstructions and GH constructions of Grazzi et al. [9]. For parity, restrict to inputs 1^t. If every transition matrix relevant to the recurrence has only real nonnegative eigenvalues, then the Jordan formula expresses each coordinate of $A(1)^t$ as a finite sum of terms

$$p(t)\lambda^t, \qquad \lambda \in \mathbb{R}_{\geq 0},$$

with polynomial p. Such terms do not oscillate between two separated finite-precision values. After finite-precision rounding, the hidden state and decoded output become eventually constant along 1^t. PARITY alternates on 1^t, so some layer must contain an eigenvalue outside $\mathbb{R}_{\geq 0}$.

For modulo m counting with m not a power of two, real eigenvalues can give at most the real sign alternation needed for period 2 after the same finite-precision stabilization. The multilayer proof applies this argument inductively: at layer i, the relevant subsequences involve products of 2^{i-1} transition matrices. To obtain periods other than powers of two, one of these products must have a nonreal eigenvalue.

For the positive construction, a generalized Householder factor has the form $I - \beta vv^\top$. Products with eigenvalue range $[-1, 1]$ have operator norm at most one. By the singular value decomposition and the Cartan–Dieudonne theorem, enough such factors represent every contraction, and orthogonal or permutation matrices require fewer reflection factors. Finite automata whose transition monoid is a group can therefore be implemented by representing transitions as permutations or rotations. In the source statement, the transition monoid embeds in a subgroup of S_n, and each one-symbol transition moving at most k points is written as at most $k-1$ swaps, hence as an element of $M^n_{k-1}(\{-1, 1\})$. Finally, Krohn–Rhodes decomposes a general finite automaton into a cascade of permutation and reset components; permutation components are handled by Householder products, and reset components use zero transitions plus the additive term B. □

A.4.12 AUSSM expressivity

AUSSM expressivity construction of Karuvally et al. [24]. If $A(u)$ is skew-symmetric, then its eigenvalues lie on the imaginary axis. Thus the discrete transition

$$\Phi(u) = \exp(\Delta A(u))$$

has eigenvalues on the complex unit circle. The input can therefore modulate rotation frequency without introducing contraction or explosion in the ideal model.

Modulo counting is the root-of-unity construction under the source's logarithmically bounded precision assumption. For counting ones modulo k, take scalar complex state $h_0 = 1$, set $B(0) = B(1) = 0$, $A(0) = 1$, and

$$A(1) = \exp(2\pi i/k).$$

After reading t ones, $h_t = \exp(2\pi it/k)$, so the state identifies the residue class modulo k as long as the k roots remain distinguishable under the assumed precision scaling.

For a cyclic group automaton, choose a generator symbol a whose transition has order k. If another symbol b acts as the n_bth power of the same cycle, set

$$A(b) = \exp(2\pi i n_b/k).$$

Then the hidden state is isomorphic to the current automaton state through the map from kth roots of unity to $\mathbb{Z}/k\mathbb{Z}$.

Mamba layers supply set-reset automata for star-free components, while AUSSM layers supply cyclic group automata. A hybrid stack can implement cascade products by padding layers that act as identities when a component is not active. Krohn–Rhodes decomposes solvable regular languages into cascades whose group factors are cyclic groups of prime order together with reset components. Combining the two layer types recognizes all such solvable languages. □

A.5 Proofs: Limitations

A.5.1 Copying separation

Copying separation of Jelassi et al. [22]. For the Transformer upper bound, choose n local heads that form keys and queries from length-n windows. If the input string has no repeated n-gram, then the current generated suffix identifies a unique earlier position in the prompt. A second attention operation retrieves the token that followed that earlier n-gram, which is exactly the next copied token. Thus the copy algorithm fails only on inputs with an n-gram collision, giving

$$\mathrm{err}_{\mathcal{D}_L}(T) < p_{n\text{-gram}}(\mathcal{D}_L).$$

For the uniform distribution, a union bound over pairs of windows gives $p_{n\text{-gram}} < L^2 D^{-n}$.

For the GSSM lower bound, after the prompt has been read, the entire future generation is a deterministic function of the final state $s \in S$. Hence the model can correctly output at most $|S|$ different length-L strings, one per state, among the D^L uniformly possible strings. Therefore

$$\Pr[\text{correct}] \le |S| D^{-L}, \qquad \mathrm{err}_{\mathcal{D}_L}(H) > 1 - |S| D^{-L}.$$

Writing $\mathrm{mem}(S) = \log |S|$ gives the stated memory threshold. □

A.5.2 Log-precision SSMs in TC^0

Circuit simulation of log-precision SSMs in Merrill et al. [29]. The generalized linear SSM convolutional form is

$$h_i = \sum_{j=1}^{i} \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j x_j, \qquad y_i = C_i h_i + D_i x_i.$$

Thus the circuit task is to compute all interval products of transitions, multiply by tokenwise terms, and perform iterated sums.

The paper proves a criterion: if each interval product $\prod_{k=j}^{i} \bar{A}_k$ is computable in L-uniform TC^0, and if $\bar{A}_i, \bar{B}_i, C_i, D_i$ are tokenwise TC^0-computable, then the whole convolutional form is in L-uniform TC^0. Non-gated SSMs satisfy the product condition because products reduce to powers of one fixed matrix. Diagonal SSMs satisfy it because interval products reduce to coordinatewise iterated scalar products. Simultaneously diagonalizable SSMs satisfy it after pulling the fixed change of basis W outside the diagonal products:

$$\prod_k W \,\mathrm{diag}(a_k) W^{-1} = W \left(\prod_k \mathrm{diag}(a_k) \right) W^{-1}.$$

S4 falls under the non-gated case, and S6/Mamba falls under the diagonal case in the formalization of the paper.

If such a model solved an NC^1-hard state-tracking problem, such as the S_5 word problem, then that problem would lie in TC^0. Assuming $\mathrm{TC}^0 \neq \mathrm{NC}^1$, this is impossible. □

A.5.3 DLOGTIME-uniform Mamba simulation

DLOGTIME-uniform Mamba simulation of Chen et al. [3]. The proof is by componentwise circuit simulation over p-bit floating point numbers with all dimensions bounded by $\mathrm{poly}(n)$. The paper first places standard floating-point arithmetic, matrix multiplication, iterated addition, iterated multiplication, exponentials, logarithms, Softplus, and SiLU inside DLOGTIME-uniform threshold circuits of constant depth and polynomial size.

A Selective SSM block is then a composition of selection functions, discretization, and either recurrent or convolutional SSM computation. Each component has a constant-depth threshold circuit, so their composition has constant depth. A Mamba block further composes input projections, a one-dimensional convolution, SiLU, the Selective SSM map, a Hadamard product, and an output projection:

$$M(X) = O\big((\mathrm{SSM}_{\mathrm{select}} \circ Z \circ C \circ L)(X) \otimes (Z \circ L)(X)\big).$$

Each named component is already in the circuit basis, so the whole block is in DLOGTIME-uniform TC^0.

The hardness statement is a containment argument. Arithmetic formula evaluation, Boolean formula value, width-5 permutation branching programs, and related permutation-composition problems are NC^1-hard or complete in the source formalization. If constant-depth Selective SSMs or Mamba blocks with polynomial precision solved these tasks, then the tasks would be in TC^0. Under $\mathrm{TC}^0 \neq \mathrm{NC}^1$, they cannot. □

A.5.4 Theorem 7.6: canonical automata and finite datatype

This follows Theorems 1–2 of Dankowiakowski and Ronca [6].

Proof. For an η-finite system, quotient the state, input, and output spaces by their path-connected compact components. Continuity of the transition map implies that a whole state component and a whole input component are mapped into one state component. Thus the quotient transition is well-defined and finite:

$$[x]_\eta, [u]_\eta \longmapsto [f(x,u)]_\eta.$$

The quotient output is well-defined for the same reason. This finite quotient is the canonical automaton, and the original system and canonical automaton produce the same decoded output sequence. Hence the implemented functions are regular.

For finite datatype implementation, strong ε-robustness supplies a margin: each state component contains an ε-ball, and every transition image has an ε-ball contained in the state space. Pick a finite datatype fine enough that every component has a representative and floating-point evaluation of f is within less than ε of the exact transition. Rounding cannot leave the intended state component, so the finite datatype system follows the same canonical automaton. □

A.5.5 Theorem 7.7: robust LRD and Mamba FLIP-FLOP obstruction

This follows Theorems 3–4 and 7 of Dankowiakowski and Ronca [6].

Proof. For an LRD, write the transition under a fixed input u as

$$x \mapsto A(u)x + B(u).$$

If u induces the identity on the canonical semiautomaton, then iterating u must keep each canonical state component fixed. The metric-automata obstruction is the source's contraction lemma for robust LRDs. Let $E_1, \ldots, E_r$ be the separated η-components representing canonical states. Strong robustness gives a positive transition margin around each $F(E_i)$, where $F(x) = A(u)x + B(u)$. Under repeated application of the same input, the affine iterates F^n cannot preserve two separated robust components indefinitely: stable directions contract, while a genuinely noncontracting affine direction would either leave the compact state space or generate a continuum of distinguishable η-components, contradicting η-finiteness. Hence, for sufficiently large n, all trajectories under the constant input are contained in one canonical component. This contradicts an identity transition on a canonical automaton with more than one state.

The FLIP-FLOP automaton has an identity instruction, together with set and reset instructions that preserve two distinct memory states. A robust η-finite LRD cascade inherits the contraction/non-identity obstruction componentwise, so it cannot implement FLIP-FLOP.

For Mamba in the source's finite/robust parameterization, the recurrent part is a cascade built from finite-context convolutions and LRD-like positive gated recurrences. The Mamba gate class cannot supply a robust identity transition on two separated canonical memory states. Applying the LRD obstruction to the cascade yields the stated FLIP-FLOP impossibility. □

A.5.6 Theorem 7.8: aperiodicity and star-free boundary

This is Theorem 6 of Dankowiakowski and Ronca [6].

Proof. Fix an input component and let A be the corresponding LRD transition matrix. All eigenvalues of A are nonnegative. Put A in Jordan normal form. For each Jordan block, either the eigenvalue has magnitude less than one, in which case iterates converge, or the block is neutral in a way that cannot produce a nontrivial cycle under

the η-finite robust quotient. Therefore, under a constant eventually η-convergent input stream, the state sequence is eventually contained in one state η-component. This is the source's metric definition of aperiodicity.

The canonical semiautomaton of an η-finite aperiodic dynamics is group-free. By algebraic automata theory, group-free finite automata recognize exactly the star-free regular languages. A finite cascade of aperiodic dynamics is aperiodic, so finite-precision Mamba-style cascades whose recurrent gates satisfy the same nonnegative-eigenvalue hypothesis have only star-free regular recognition power in this setting. □

A.5.7 Theorem 7.9: geometric Mamba results

This follows Theorems 8–9 of Dankowiakowski and Ronca [6].

Proof. For the positive result, the paper constructs Mamba-parameterized dynamics as geometrically constrained systems that realize the primitive components used in group-free automata, including FLIP-FLOP-like memory inside separated convex regions rather than finite robust η-components. Since star-free languages are recognized by group-free automata, and group-free automata decompose into cascades of such primitive components, Mamba GCS cascades can recognize all star-free languages.

For the limitation, replace η-components by components of a convex-regular covering. If an η-finite LRD has nonnegative-eigenvalue transition gates, the same Jordan-form convergence argument as in Theorem 7.8 implies that state trajectories under eventually constant input are eventually trapped in one covering component. Thus the dynamics are aperiodic with respect to the covering. An alternating function changes value infinitely often on some constant-symbol input sequence, but a continuous readout on one eventual covering component is eventually constant. Hence alternating functions are not implementable in that convex-regular GCS setting. □

A.5.8 Unified parity obstruction

Unified parity obstruction of Khavari et al. [25]. The construction first notes a contrast: a time-invariant S4D layer can solve modular counting on the fixed input 1^t by a root-of-unity transition. This does not solve parity on arbitrary binary strings, because the needed oscillation is not input-dependent.

For nonnegative diagonal Mamba-like layers, consider repeated blocks

$$0^k 1^m 0^k 1^m \cdots$$

with m odd. Over one block, the product of nonnegative diagonal transition matrices is again nonnegative diagonal. The block-level recurrence therefore has the same finite-precision collapse behavior as the nonnegative parity lower bound, while the correct parity flips after each block.

For a stack containing S4D and Mamba layers, choose a block length W that is a common period for the rational S4D phases used in the proof. At the W-block scale, the S4D transitions become real nonnegative, and Mamba transitions are already nonnegative. Skip connections and learned initial states only add finitely many block-level affine terms. Induction over layers then shows that the block-sampled states become stationary at finite precision, whereas parity continues to alternate. Hence the diagonal S4D+Mamba stack cannot solve parity at arbitrary lengths. □

A.6 Proofs: Stability and Generalization

A.6.1 Auxiliary concentration and complexity estimates

The proofs in Sections 8 and 9 repeatedly use the following standard estimates. They are recorded here to make the later arguments readable without changing the source theorem statements.

Symmetrization and bounded losses. Let $Z_1, \dots, Z_n$ be i.i.d. with law P, let P_n be the empirical law, and let $\mathcal{F}$ be a real-valued class for which the following expectations are finite. If $\sigma_1, \dots, \sigma_n$ are independent Rademacher signs, then

$$\mathbb{E} \sup_{f \in \mathcal{F}} |Pf - P_n f| \le 2\mathbb{E}_{Z,\sigma} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right|.$$

Indeed, with an independent ghost sample $Z'_1, \dots, Z'_n$,

$$\mathbb{E}\sup_f(Pf - P_n f) \le \mathbb{E}\sup_f(P'_n f - P_n f) = \mathbb{E}\sup_f \frac{1}{n}\sum_i \sigma_i(f(Z'_i) - f(Z_i)),$$

and the triangle inequality gives the factor 2; the reverse deviation is identical. If $0 \le f \le K_\ell$, McDiarmid or Hoeffding applied to the supremum adds, with probability at least $1-\delta$,

$$K_\ell\sqrt{\frac{2\log(4/\delta)}{n}}$$

to the expected supremum. This is the conversion used in the length-independent PAC estimate below.

Covering to Rademacher complexity. For a B-bounded class $\mathcal{F}$ on a sample S of size m, Dudley's entropy integral gives

$$\mathrm{Rad}_S(\mathcal{F}) \le \inf_{\alpha>0}\left\{4\alpha + \frac{12}{\sqrt{m}}\int_\alpha^B \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}, L_2(S))}\, d\varepsilon\right\}.$$

If a Lipschitz loss ℓ with Lipschitz constant $\mathfrak{l}_\ell$ and range bound $\mathfrak{c}_\ell$ is composed with a hypothesis class whose covering entropy is summarized by a source capacity $\mathcal{C}$, optimizing the preceding integral at the source choice of scale yields the displayed form

$$\frac{12\mathfrak{l}_\ell\mathcal{C}}{\sqrt{m}}\left(1 + \log\frac{\mathfrak{c}_\ell\sqrt{m}}{3\mathcal{C}}\right) + 3\mathfrak{c}_\ell\sqrt{\frac{\log(2/\delta)}{2m}}.$$

The constants are the standard Dudley and bounded-difference constants; the model-specific work is to prove the capacity bound for $\mathcal{F}$.

Products of transition matrices. For two products of equal length,

$$A_q\cdots A_1 - \widehat{A}_q\cdots\widehat{A}_1 = \sum_{\ell=1}^{q} A_q\cdots A_{\ell+1}(A_\ell - \widehat{A}_\ell)\widehat{A}_{\ell-1}\cdots\widehat{A}_1.$$

Thus, if all factors have norm at most ρ_A and $\|A_\ell - \widehat{A}_\ell\| \le \varepsilon_A$, then

$$\|A_q\cdots A_1 - \widehat{A}_q\cdots\widehat{A}_1\| \le q\rho_A^{q-1}\varepsilon_A.$$

When the product appears inside a causal sum over all lengths $q = 0, \dots, T-1$, the sharper geometric sums in the cited selective-SSM covering result give accumulated coefficients of the form

$$\frac{\rho_A(1-\rho_A^T)}{(1-\rho_A)^2} - \frac{T\rho_A^T}{1-\rho_A},$$

with the continuous extension at $\rho_A = 1$.

Gaussian and Gaussian-chaos concentration. If $X_1, \dots, X_m$ are independent mean-zero sub-Gaussian variables with sub-Gaussian scale at most K, then for an absolute constant $c > 0$,

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^m X_i\right| > t\right) \le 2\exp\left(-\frac{cmt^2}{K^2}\right).$$

A union bound makes this uniform over polynomially many fixed directions. For a degree-q polynomial $Q(g)$ of a standard Gaussian vector with $\|Q - \mathbb{E}Q\|_{\psi_{2/q}} \le K$, the corresponding sub-Weibull Bernstein bound is

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^m (Q(g_i) - \mathbb{E}Q)\right| > t\right) \le 2\exp\left(-cm\min\{(t/K)^2, (t/K)^{2/q}\}\right).$$

Gaussian hypercontractivity supplies the required sub-Weibull norm for the fixed low-degree Hermite polynomials used in the ICL feature-learning proof.

Packing lower-bound template. Let Θ_0 be a finite set of parameters whose induced targets are 2ε-separated in the prediction metric. If an algorithm with hidden state space $\mathcal{H}$ attains error at most ε uniformly over Θ_0, then two parameters in Θ_0 cannot lead to the same distinguishable hidden state on the separating prompts. Hence $|\mathcal{H}| \geq |\Theta_0|$. For a d-dimensional recurrent state stored at precision $\mathtt{p}$, $|\mathcal{H}| \leq 2^{d\mathtt{p}}$, so any packing of size M implies

$$d\mathtt{p} \geq \log_2 M.$$

The Markov-chain lower bound below uses $M = \exp(2^k(1-3\varepsilon)\log(1/\varepsilon))$.

A.6.2 StableSSM memory theorem

Proof of Wang and Li [42], Theorem 3.3. The source memory function $\mathcal{M}(\mathbf{H})(t)$ is expressed through the derivative response to Heaviside inputs. Consider a one-layer SSM approximant

$$\dot{h}_{m,t} = \Lambda_m h_{m,t} + U_m x_t + b_m, \qquad \widehat{y}_{m,t} = c_m^\top \sigma(h_{m,t}),$$

and a Heaviside input $x_t = x_0 \mathbf{1}_{\{t \geq 0\}}$. Put $v_{m,t} = dh_{m,t}/dt$. For $t \geq 0$, differentiating the state equation gives

$$\dot{v}_{m,t} = \Lambda_m v_{m,t}, \qquad v_{m,0} = U_m x_0 + b_m,$$

where the initial hidden state is zero before the jump. The source first uses stable approximation to perturb the recurrent matrix by a positive shift inside the allowed radius. Since the perturbed systems remain asymptotically stable, the unperturbed diagonal matrices satisfy the uniform spectral bound

$$\sup_m \max_i \Re\lambda_i(\Lambda_m) \leq -\beta_0.$$

Therefore

$$\|v_{m,t}\|_2 = \|e^{\Lambda_m t} v_{m,0}\|_2 \leq e^{-\beta_0 t} \|U_m x_0 + b_m\|_2.$$

Using the uniform parameter bound and then taking the normalization in the definition of $\mathcal{M}$,

$$\|U_m x_0 + b_m\|_2 \leq (d\|x_0\|_\infty + 1)\theta_{\max},$$

and so

$$\mathcal{M}(\widehat{\mathbf{H}}_m)(t) = \sup_{x_0} \frac{1}{\|x_0\|_\infty + 1} \left| \frac{d\widehat{y}_{m,t}}{dt} \right| \leq (d+1) L_0 \theta_{\max}^2 e^{-\beta_0 t}.$$

Here the displayed estimate is valid with rate β_0 for each approximant; the source's transfer lemma passes from uniformly exponentially decaying approximants to the target and yields the stated bound for every $0 < \beta < \beta_0$.

For ℓ layers, the derivative signal passes through a product of ℓ stable transition kernels. Convolution of ℓ exponentially decaying kernels produces a polynomial prefactor of degree at most $\ell - 1$ times the same exponential decay:

$$\mathcal{M}(\mathbf{H})(t) \leq (d+1) L_0^\ell \theta_{\max}^{\ell+1} P(t) e^{-\beta t}.$$

□

A.6.3 Stable reparameterization and gradient scale

Proof of Wang and Li [42], Theorems 3.5 and 3.6. For linear time-homogeneous functionals, the approximation problem reduces to approximating an L^1 memory kernel by exponential sums

$$\rho(t) \approx \widehat{\rho}_m(t) = \sum_{i=1}^m c_i e^{\lambda_i t}, \qquad \lambda_i < 0.$$

If $\lambda_i = f(w_i)$ and f is stable, then perturbing w_i by at most β changes the corresponding exponential kernel by an L^1 amount controlled uniformly by the defining inequality

$$|f(w_i)| \sup_{|\widetilde{w}_i - w_i| \leq \beta} \int_0^\infty |e^{f(\widetilde{w}_i)t} - e^{f(w_i)t}|\, dt \leq g(\beta).$$

After summing the finitely many exponential modes and using the linearity of the readout, the perturbation error of the approximating linear RNN is bounded by the approximation error plus a term controlled by $g(\beta)$. This is the source's simplified perturbation setting: the stability perturbation is placed on the recurrent weights. The approximation error goes to zero with m, and $g(\beta) \to 0$ as $\beta \downarrow 0$. Hence the approximation is stable in the stated sense.

For the gradient statement, differentiate the loss through $\lambda = f(w)$. The derivative of the exponential mode satisfies the scale estimate

$$\left\| \frac{\partial}{\partial \lambda} e^{\lambda t} \right\|_{L^1(0,\infty)} = \int_0^\infty t e^{\lambda t}\, dt = \frac{1}{|\lambda|^2}, \qquad \lambda < 0.$$

The chain rule therefore gives

$$\left| \frac{\partial \,\mathrm{Loss}}{\partial w} \right| \le C_{\mathbf{H},\widehat{\mathbf{H}}_m} \frac{|f'(w)|}{f(w)^2}.$$

In discrete time the corresponding geometric derivative has scale $(1 - f(w))^{-2}$, giving the stated discrete bound.

Solving the source "balanced gradient" criterion

$$\frac{|f'(w)|}{f(w)^2} = L|w|$$

with negative continuous-time eigenvalues gives

$$\frac{d}{dw}\left(-\frac{1}{f(w)}\right) = 2aw, \qquad f(w) = -\frac{1}{aw^2 + b}.$$

The same calculation with the discrete stability boundary at 1 gives $f(w) = 1 - \frac{1}{aw^2+b}$. □

A.6.4 Data-dependent LTI SSM bound

Proof of Liu and Li [28], Theorem 1 and Proposition 1. Write

$$x(s) = \mu(s) + \sqrt{K(s,s)}\,\widetilde{x}(s).$$

For a fixed θ,

$$\left| \int_0^T \rho_\theta(T-s) x(s)\, ds \right| \le \int_0^T |\rho_\theta(T-s)| \sqrt{K(s,s)} |\widetilde{x}(s)|\, ds + \left| \int_0^T \rho_\theta(T-s)\mu(s)\, ds \right|.$$

Taking the supremum over $\theta \in \Theta$ introduces exactly the source quantity $\psi(\Theta)$, up to the random factor $\sup_{s\in[0,T]} |\widetilde{x}(s)|$.

The empirical-process proof applies the symmetrization estimate recorded in Appendix A.6.1:

$$\mathbb{E} \sup_{\theta\in\Theta} |R_x(\theta) - R_n(\theta)| \lesssim \mathbb{E}_{\sigma,x} \sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \int_0^T \rho_\theta(T-s) x_i(s)\, ds \right|.$$

The Holder condition gives an ε-net over $[0,T]$ for the normalized paths. Holder continuity controls the oscillation between net points, while the pointwise sub-Gaussian tail, the sub-Gaussian sample-mean estimate, and a union bound control the maximum over the net. Optimizing the mesh size yields the $\log^{3/2}(Tn/\delta)$ factor. Passing from the linear functional class to the loss class by the assumed Lipschitz/quadratic loss estimate produces the square $(\psi(\Theta)+1)^2$, giving

$$\sup_{\theta\in\Theta} |R_x(\theta) - R_n(\theta)| \le (\psi(\Theta)+1)^2 O\left(\frac{\log^{3/2}(Tn/\delta)}{\sqrt{n}} \right).$$

For the initialization proposition, rescaling C to $\widetilde{C} = C/\sqrt{\tau(\theta)}$ rescales the whole kernel:

$$\rho_{\widetilde{\theta}}(s) = \frac{\rho_\theta(s)}{\sqrt{\tau(\theta)}}.$$

Write $x(s) = \mu(s) + \sqrt{K(s,s)}\,\widetilde{x}(s)$ and set

$$A_\theta = \int_0^T |\rho_\theta(T-s)|\sqrt{K(s,s)}\,ds, \qquad B_\theta = \left|\int_0^T \rho_\theta(T-s)\mu(s)\,ds\right|.$$

Since $\sqrt{\tau(\theta)} = A_\theta + B_\theta$,

$$\begin{aligned}\left|\int_0^T \rho_{\widetilde{\theta}}(T-s)x(s)\,ds\right| &\le \frac{A_\theta}{A_\theta + B_\theta}\sup_{s\in[0,T]}|\widetilde{x}(s)| + \frac{B_\theta}{A_\theta + B_\theta} \\ &\le \sup_{s\in[0,T]}|\widetilde{x}(s)| + 1.\end{aligned}$$

Applying the Holder/sub-Gaussian supremum estimate to $\sup_{s\in[0,T]}|\widetilde{x}(s)|$ gives

$$\mathbb{E}_x\left|\int_0^T \rho_{\widetilde{\theta}}(T-s)x(s)\,ds\right| \le O(\sqrt{\log T}),$$

with constants depending only on the process constants. □

A.6.5 Length-independent deep SSM PAC bound

Proof of Rácz et al. [33]. The control-theoretic input is that an internally stable discrete-time SSM has sequence operators whose induced norms are bounded uniformly over the horizon T. In the source notation,

$$\sup_T \|S_{\Sigma,T}\|_{2,\infty} \le \|\Sigma\|_1, \qquad \sup_T \|S_{\Sigma,T}\|_{\infty,\infty} \le \|\Sigma\|_2.$$

Thus an SSM layer is Rademacher contractive with multiplicative constant controlled by a system norm, not by the sequence length.

The Rademacher-contraction composition lemma is a nesting argument. If Φ_1 maps X_1 to X_2 and Φ_2 maps X_2 to X_3, apply the definition first to the outer class Φ_2 with the admissible set $\{\varphi_1(u_i)\}_i$, and then apply it to Φ_1:

$$\begin{aligned}&\mathbb{E}_\sigma \sup_{\varphi_2,\varphi_1,Z}\left\|\frac{1}{M}\sum_i \sigma_i\varphi_2(\varphi_1(u_i))\right\|_{\mathcal{X}_3} \\ &\quad\le \mu_2\mathbb{E}_\sigma \sup_{\varphi_1,Z}\left\|\frac{1}{M}\sum_i \sigma_i\varphi_1(u_i)\right\|_{\mathcal{X}_2} + \frac{c_2}{\sqrt{M}} \\ &\quad\le \mu_2\left(\mu_1\mathbb{E}_\sigma \sup_{Z}\left\|\frac{1}{M}\sum_i \sigma_i u_i\right\|_{\mathcal{X}_1} + \frac{c_1}{\sqrt{M}}\right) + \frac{c_2}{\sqrt{M}}.\end{aligned}$$

This gives $(\mu_1\mu_2, \mu_2 c_1 + c_2)$.

The paper proves that the encoder, decoder, pooling layer, stable SSM layer, and allowed MLP/GLU blocks are all Rademacher contractive on the relevant balls, with constants tabulated in the source. Iterating the composition lemma through the architecture gives the displayed constants μ, c and the radius recursion r_i. The input bound gives

$$\mathbb{E}_\sigma\left\|\frac{1}{M}\sum_i \sigma_i u_i\right\| \le \frac{K_u}{\sqrt{M}}.$$

Finally, the symmetrization and bounded-loss conversion in Appendix A.6.1 bound the empirical-process term by the Rademacher complexity of the loss class. The contraction step contributes $(\mu K_u L_\ell + cL_\ell)/\sqrt{M}$, and Hoeffding's inequality for the bounded loss supplies the confidence term, yielding

$$\mathcal{L}(f) - \mathcal{L}^S_{\text{emp}}(f) \le \frac{\mu K_u L_\ell + cL_\ell}{\sqrt{M}} + K_\ell\sqrt{\frac{2\log(4/\delta)}{M}}.$$

Since the component constants use stable system norms bounded uniformly in T, the final bound is length-independent. □

A.6.6 Selective SSM covering bounds

Proof of Honarpisheh et al. [19]. Unroll the selective recurrence. Each output at time T is a sum over past tokens whose coefficients contain products of input-dependent transitions. This exposes a causal linear-attention-like structure in the parameters W_B, W_C, q, w, but with an additional transition product depending on A_c, p, q.

The transition product is controlled by the spectral abscissa. Since

$$A[t] = \exp(\Delta[t]A_c), \qquad \Delta[t] = \mathrm{softplus}(p + q^\top u[t]),$$

bounded inputs and bounded q give

$$0 < \Delta[t] \le \mathfrak{M}_\Delta = \log(1 + e^{p+\mathfrak{B}_q\mathfrak{B}_u}).$$

Gelfand's formula applied to e^{A_c} gives, for any small $\eta > 0$,

$$\|A^t\|_2 \le \rho_A^t, \qquad \rho_A = (1 + e^{p-\mathfrak{B}_q\mathfrak{B}_u})^{s_A+\eta}.$$

The product-telescoping identity in Appendix A.6.1 then bounds the difference between transition products formed from A_c and from a cover point $\widehat{A}_c$; the accumulated geometric sums produce the source factor

$$S_2 = \sum_{t=0}^{T-1} t\rho_A^t = \frac{\rho_A(1-\rho_A^T)}{(1-\rho_A)^2} - \frac{T\rho_A^T}{1-\rho_A}.$$

The same finite sum gives the limiting value $T(T-1)/2$ when $\rho_A = 1$.

The remaining parameters are covered as linear function classes with bounded operator and mixed norms. The cover for A_c is combined with the covers for W_B, W_C, q, w by a Cartesian product. If the target output error is ε, choose radii $\varepsilon_A, \varepsilon_B, \varepsilon_C, \varepsilon_q, \varepsilon_w$ so the Lipschitz and telescoping errors add to at most ε; optimizing the resulting product of covering numbers gives the capacity expression in the theorem. The covering-to-Rademacher estimate in Appendix A.6.1 and the bounded Lipschitz loss assumption convert this cover into the stated generalization bound with capacity $\mathcal{C}_{\mathcal{F}_{\mathrm{SSM}}}$.

For the linear-attention specialization, use the source's constant-step normalization $\Delta[t] = 1$ together with $A_c = 0$, so $A[t] = I$. The transition-product cover disappears, and the unrolled expression is the causal linear-attention query-key-value sum. The same covering argument then yields

$$\mathcal{C}_{\mathcal{F}_{\mathrm{LA}}} = \widetilde{O}(T\mathfrak{B}_w\mathfrak{B}_B\mathfrak{B}_C\mathfrak{B}_u^3).$$

For the lower bound, restrict to a one-parameter subclass of selective SSMs with fixed step size and a scalar output weight w. On inputs in $[-1,1]^T$, the output contains a sum of transition magnitudes growing like

$$1 + (1+s_A) + \cdots + (1+s_A)^{T-1} = \frac{(1+s_A)^T - 1}{s_A}$$

when $s_A > 0$, and equals T when $s_A = 0$. The empirical Rademacher complexity of this restricted subclass is therefore at least the output scale times

$$\mathbb{E}_\sigma \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \right| = \sqrt{\frac{2}{\pi m}},$$

up to the allowed factor $\mathfrak{B}_w$. A lower bound for a subclass is a lower bound for the full class. □

A.6.7 Lyapunov stability of MambaBlock recurrences

Proof of Halloran et al. [18], Theorem 1. For the source MambaBlock recurrence,

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \qquad \bar{A}_t = \exp(\bar{\Delta}_t A),$$

the recurrent Jacobian is

$$\frac{\partial h_t}{\partial h_{t-1}} = \bar{A}_t.$$

Under the diagonal Mamba parameterization,

$$A = \operatorname{diag}(a_1, \dots, a_N), \qquad \bar{\Delta}_t = \operatorname{diag}(\delta_{t,1}, \dots, \delta_{t,N}), \qquad \delta_{t,i} \geq 0, \quad \operatorname{Re} a_i \leq 0.$$

Thus

$$\bar{A}_t = \operatorname{diag}(e^{\delta_{t,1} a_1}, \dots, e^{\delta_{t,N} a_N}),$$

and all transition factors commute. Hence

$$\prod_{t=1}^{L} \bar{A}_t = \operatorname{diag}\left(\exp\left(a_i \sum_{t=1}^{L} \delta_{t,i} \right) \right)_{i=1}^{N}.$$

Taking the spectral norm gives

$$\left\| \prod_{t=1}^{L} \bar{A}_t \right\|_2 = \max_i \exp\left(\operatorname{Re} a_i \sum_{t=1}^{L} \delta_{t,i} \right) \leq 1.$$

Therefore

$$\lambda_{\max} = \limsup_{L \to \infty} \frac{1}{L} \log \left\| \prod_{t=1}^{L} \frac{\partial h_t}{\partial h_{t-1}} \right\|_2 \leq 0.$$

Let

$$J_t = \frac{\partial h_t}{\partial h_{t-1}}.$$

For any $\eta > 0$, the definition of the limsup gives a constant C_η such that every sufficiently long product satisfies

$$\|J_{t+n} \cdots J_{t+1}\|_2 \leq C_\eta e^{n(\lambda_{\max} + \eta)}.$$

The mean-value formula for the iterated state map gives

$$F_\theta^n(z + \delta) - F_\theta^n(z) = \left(\int_0^1 DF_\theta^n(z + s\delta)\, ds \right) \delta.$$

Absorbing C_η and the local derivative bounds into the big-O constant, and taking any source envelope $\zeta \geq \lambda_{\max} + \eta$ with $\zeta \leq 0$, gives

$$\max |F_\theta^n(h_{t-1}, x_t) - F_\theta^n(h_{t-1} + \varepsilon, x_t + \varepsilon)| \in O(|\varepsilon| e^{n\zeta})$$

with $\zeta = \lambda_{\max} \leq 0$ or any source upper envelope $\zeta \geq \lambda_{\max}$ satisfying $\zeta \leq 0$. □

Proof of Halloran et al. [18], Theorems 2–4. For outputs, $y_t = C_t h_t$. If $\sup_t \|C_t\|_2 \leq M_C$, then

$$\|y_t - y_t'\|_2 = \|C_t(h_t - h_t')\|_2 \leq M_C \|h_t - h_t'\|_2.$$

Thus any latent-state perturbation estimate $O(|\varepsilon| e^{t\zeta})$ transfers to the output, with M_C absorbed into the big-O constant.

For the MPFT/PEFT statement, the source compares a full-precision trajectory $\{h_t, x_t\}$ with a perturbed trajectory $\{h_t', x_t'\}$ and sets

$$\varepsilon^* = \max_{0 \leq t \leq L} \max\{\|h_t - h_t'\|_\infty, \|x_t - x_t'\|_\infty\}.$$

Every state/input discrepancy vector has Euclidean norm at most $\sqrt{d}\,\varepsilon^*$ for the relevant source dimension d. Applying the Lyapunov envelope from the preceding proof to these perturbations gives

$$\max \left| F_\theta^L(h_0, x_1) - F_\theta^L(h_0', x_1') \right| \in O(\varepsilon^* e^{L\zeta}), \qquad \zeta \leq 0,$$

where the fixed factor $\sqrt{d}$ is absorbed into the constant. Multiplying by the bounded output map C_L gives

$$\max |y_L - y_L'| \in O(\varepsilon^* e^{L\zeta}).$$

This is the source perturbative conclusion for deviations represented by mixed-precision and LoRA/PEFT changes inside the MambaBlock. □

A.6.8 Training dynamics for simplified Mamba blocks

Proof of Shandirasegaran et al. [37]. The analyzed model is a simplified Mamba block with trainable output weights and trainable gate vector w_Δ, while the B and C projections are kept at initialization in the theorem statements. Write the gate update as

$$w_\Delta^{(t+1)} = w_\Delta^{(t)} - \eta \nabla_{w_\Delta} \widehat{\mathrm{Loss}}^{(t)}.$$

Expanding this gradient in the orthonormal feature basis $o_+, o_-, o_3, \ldots, o_d$ separates contributions from class-relevant, confusing, and irrelevant directions. The Gaussian sample-mean estimates in Appendix A.6.1 make the empirical gradient coefficients match their population signs throughout training. The argument is carried out under the source structured data model: tokens are noisy copies of an orthonormal feature family, and the training set is balanced between labels.

For majority-voting data, concentration of the Gaussian initialization and the sample average implies that the gradient component along o_+ and o_- has a positive drift proportional to the squared majority gap $(\alpha_r L - \alpha_c L)^2/L^2$. Summing the drift over T_{gd} gradient steps with learning rate η gives

$$\langle w_\Delta^{(T_{\mathrm{gd}})}, o_\pm \rangle \geq \frac{\eta T_{\mathrm{gd}}}{8L^2} \Theta((\alpha_r L - \alpha_c L)^2).$$

The irrelevant directions have only fluctuation-scale contributions after the same concentration step, yielding the $\widetilde{O}(1/\operatorname{poly}(d))$ bound. Once the gate has this alignment, the recurrent aggregation gives a margin with the correct sign on the majority-voting distribution. The stated sample lower bound ensures the empirical gradients have the population signs, and the stated iteration count is exactly the time needed for the margin to dominate noise.

The majority-voting generalization theorem uses the preceding alignment as its margin certificate. The width assumption gives enough initialized output features to separate the two labels with high probability; the noise condition $\tau < O(1/d)$ keeps token perturbations below the margin. The sample lower bound makes all empirical gradient signs match their population signs throughout the T_{gd} iterations, and the iteration count is chosen so the aligned gate contribution dominates the confusing-token contribution. Thus the source target-error functional vanishes.

For locality-structured data, the class-relevant features need not have a consistent positive drift because useful and confusing counts may be comparable. The proof instead shows that irrelevant directions acquire a negative drift determined by the locality gap

$$(1/2)^{\Delta L_{o_+}^+} - (1/2)^{\Delta L_{o_+}^-}.$$

Summing this drift gives the displayed upper bound for $\langle w_\Delta^{(T_{\mathrm{gd}})}, o_j \rangle$, while the o_+ and o_- components remain within the stated fluctuation scale. The recurrent part then uses locality to separate class-relevant patterns from confusing ones.

The locality-structured generalization theorem combines this negative irrelevant-direction drift with the recurrent locality separation. The sample lower bound is the concentration condition needed for the locality-gradient signs, while the iteration bound is the time needed for the locality margin, proportional to

$$(1/2)^{\Delta L_{o_+}^+} - (1/2)^{\Delta L_{o_+}^-},$$

to dominate fluctuations. Under these two conditions the source target-error functional again vanishes with the stated high probability. □

A.7 Proofs: In-Context Learning and Algorithmic Behavior

A.7.1 SSMs implementing gradient descent

Construction in Sushma et al. [39]. For scalar linear regression, the unscaled gradient after t examples is

$$g_{w_0}(D_{1:t}) = \sum_{i=1}^{t} (w_0^\top x_i - y_i) x_i.$$

It satisfies the recurrence

$$g_{w_0}(D_{1:t}) = g_{w_0}(D_{1:t-1}) + (w_0^\top x_t - y_t) x_t.$$

When $w_0 = 0$, the one-step gradient-descent prediction on x_{t+1} is

$$\widehat{y}_{t+1} = -\frac{\eta}{t} g_{w_0}(D_{1:t})^\top x_{t+1}.$$

Thus the recurrent state can be chosen to equal the accumulated gradient statistic. One writes this as

$$z_t = I z_{t-1} + \Psi c_t, \qquad o_t = \beta z_t^\top \Theta c_t,$$

where, in the zero-initialized construction stated in the paper, $c_t = [y_t x_t, x_{t+1}]$. The matrices Ψ, Θ select the product coordinate $y_t x_t$ and pair the state with x_{t+1}; the source implements the required sign and learning-rate normalization in the readout. The multiplication in the output gate computes the inner product between the accumulated gradient and the next input.

For vector-valued regression, let Z_t be a matrix state. With a sliding window containing (x_t, y_t, x_{t+1}), the update

$$Z_t = Z_{t-1} + y_t x_t^\top$$

accumulates the sufficient statistic for the zero-initialized linear model, and the output

$$o_t = -\frac{\eta}{N} Z_t x_{t+1}$$

is the one-step GD prediction. The paper realizes $y_t x_t^\top$ as $C_t Q C_t^\top$ for a fixed selector Q applied to the context window C_t.

For multiple GD steps, later layers require the nonzero initialization $W_{\ell-1}$ produced by previous steps. The gradient decomposes into a $y_t x_t^\top$ accumulator and an $x_t x_t^\top$ accumulator:

$$\Delta_\ell W = -\frac{\eta}{N}\left(W_{\ell-1}^\top \widetilde{Z}_t^{(\ell)} - Z_t^{(\ell)}\right).$$

The construction uses two recurrent heads per layer to maintain these two statistics. Repeating the construction layer by layer yields ℓ steps of gradient descent.

For the nonlinear variant, the recurrent state first stores the same gradient-like linear statistic. An MLP then maps this statistic to

$$\sum_{i=1}^{t} (g(w_0^\top x_i) - y_i) g'(w_0^\top x_i) x_i,$$

which is the nonlinear least-squares gradient. The final multiplicative readout pairs the transformed statistic with x_{t+1}. □

A.7.2 Mamba and Laplace smoothing for Markov chains

Bayes estimator and MambaZero construction in Bondaschi et al. [2]. For a fixed length-k context i_1^k, write $p = P_{i_1^k}(1)$. Under the Dir(β) prior, after seeing n_1 ones and n_0 zeros following that context, the posterior density is proportional to

$$p^{n_1+\beta-1}(1-p)^{n_0+\beta-1}.$$

The Bayes-optimal cross-entropy predictor is the posterior predictive probability

$$\mathbb{E}[p \mid x_1^t] = \frac{\int_0^1 p^{n_1+\beta}(1-p)^{n_0+\beta-1}\,dp}{\int_0^1 p^{n_1+\beta-1}(1-p)^{n_0+\beta-1}\,dp} = \frac{n_1+\beta}{n_0+n_1+2\beta}.$$

The corresponding probability for zero is $(n_0+\beta)/(n+2\beta)$.

For the MambaZero representation theorem, set the transition factor $a_t = 1$ so the hidden state does not forget counts. With convolution window $w = 2$, the input-selective terms can depend on the last transition $x_{t-1} \to x_t$. Unrolling the state gives

$$H_t = \widetilde{x}_0 b_0^\top + \sum_{i,j\in\{0,1\}} n_{ij}\, \widetilde{x}^{(ij)} (b^{(ij)})^\top,$$

where n_{ij} is the number of observed transitions $i \to j$. The output projection c_t selects the row associated with the current symbol x_t. The parameter choice makes the two count vectors for $n_{x_t 0}$ and $n_{x_t 1}$ linearly independent and arranges the non-count terms to contribute (β, β) up to an arbitrarily small error. After the final L^1 normalization, the logits therefore implement

$$\left(\frac{n_{x_t 0} + \beta}{n_{x_t 0} + n_{x_t 1} + 2\beta}, \frac{n_{x_t 1} + \beta}{n_{x_t 0} + n_{x_t 1} + 2\beta} \right)$$

within the prescribed KL tolerance. □

A.7.3 Finite-precision recurrent lower bound for add-β estimation

Lower bound of Bondaschi et al. [2]. For an order-k binary Markov chain, there are 2^k contexts. Under the Dirichlet prior, the conditional distributions associated with these contexts are independent. To approximate the add-1 posterior predictive distribution uniformly, a recurrent model must distinguish enough different posterior predictive vectors for all 2^k contexts.

With hidden dimension d and precision $\mathtt{p}$, the recurrent state has at most $2^{d\mathtt{p}}$ distinguishable values. The packing template in Appendix A.6.1 reduces the lower bound to constructing many Markov kernels whose context-conditional distributions are separated by more than the allowed ε error. A predictor satisfying

$$\|\widehat{p}_t - \mathbb{P}_1^{(k)}(\cdot \mid x_1^t)\|_\infty \le \varepsilon$$

must assign distinct enough hidden representations to this separated family. The packing size of the family gives the lower bound

$$2^{d\mathtt{p}} \ge \exp\left(2^k(1-3\varepsilon)\log(1/\varepsilon)\right),$$

equivalently

$$d\mathtt{p} \ge 2^k(1-3\varepsilon)\log(1/\varepsilon).$$

The argument uses only sequential compression into the hidden state, so it also applies to time-varying recurrent updates. □

A.7.4 Trained Mamba emulating online gradient descent

Proof of Jiang et al. [23]. Under the fixed choice $A = -I$, $w_\Delta = 0$, and $b_\Delta = \log(\exp((\log 2)/N) - 1)$, the recurrence has scalar decay

$$\alpha = \exp(-(\log 2)/N).$$

The selected B and C maps reduce the query output to a bilinear expression in the trainable matrices:

$$\widehat{y}_q = (1-\alpha)(Cx_q + b_C)^\top \sum_{i=0}^{N-1} \alpha^{i+1} y_{N-i}(Bx_{N-i} + y_{N-i}b + b_B).$$

Computing the population gradient over the Gaussian linear-regression task yields the matrix update system

$$B(t+1) = B(t) + \eta\beta_3 C(t) - \eta\beta_1 C(t)C(t)^\top B(t),$$

$$C(t+1) = C(t) + \eta\beta_3 B(t) - \eta\beta_1 B(t)B(t)^\top C(t) - \eta\beta_2 b(t)b(t)^\top C(t),$$

$$b(t+1) = b(t) - \eta\beta_2 C(t)C(t)^\top b(t), \qquad b_B(t) = b_C(t) = 0.$$

Decomposing $B = [b_1, \ldots, b_d]$ and $C = [c_1, \ldots, c_d]$, the proof studies the coupled dynamics of $c_i^\top b_i$, $c_i^\top b_j$, and $c_i^\top b$. The Gaussian concentration estimates in Appendix A.6.1 give nondegenerate vector norms and small cross-interactions. A fine-grained induction then maintains these properties and proves exponential convergence

$$C^\top B \to \frac{\beta_3}{\beta_1} I, \qquad C^\top b \to 0.$$

Substituting these limits into the recurrence and output formulas gives

$$(W_C^\top)_{[1:d,:]} h_l^{(d+1)} = \alpha (W_C^\top)_{[1:d,:]} h_{l-1}^{(d+1)} + (1-\alpha)\frac{\beta_3}{\beta_1} y_l x_l,$$

and hence the exponentially weighted prediction

$$\widehat{y}_q = x_q^\top \sum_{i=0}^{N-1} (1-\alpha)\alpha^{i+1} \frac{\beta_3}{\beta_1} y_{N-i} x_{N-i}.$$

The constants $\beta_1, \beta_2, \beta_3$ are Gaussian moments of the task distribution. Evaluating them gives

$$\frac{\beta_3}{\beta_1} = \frac{2(1+\alpha)}{\alpha(3(1-\alpha)d + 4 - 2\alpha)}$$

and the remaining Gaussian moment calculation gives

$$\mathcal{L}(\theta(t)) \le \frac{3d(d+1)}{2N}.$$

□

A.7.5 ICL with outliers

Proof of Li et al. [27]. Unrolling the one-layer Mamba recurrence rewrites the output on a prompt as

$$F(\Psi; P) = \sum_{i=1}^{l+1} G_{i,l+1}(w)\, y_i\, p_i^\top W_B^\top W_C p_{\text{query}},$$

where

$$G_{i,l+1}(w) = \sigma(w^\top p_i) \prod_{j=i+1}^{l+1} (1 - \sigma(w^\top p_j))$$

for $i < l+1$. Thus the model is a linear-attention score $p_i^\top W_B^\top W_C p_{\text{query}}$ multiplied by a nonlinear gate.

The training proof tracks SGD in two parts. First, gradients of W_B, W_C increase the linear-attention score for examples sharing the relevant pattern with the query, while keeping other relevant-pattern scores small. Second, the gate parameter w is analyzed in two phases. In the first phase it begins to separate outlier directions from clean directions; in the second phase the sigmoid gate makes outlier-containing examples have small gate values. The batch-size, outlier-magnitude, context-length, and iteration assumptions are chosen so the concentration bounds from Appendix A.6.1 keep these gradient signs stable over training. Combining the attention and gate estimates gives

$$\mathbb{E}_{f\in\mathcal{T}, P\sim\mathcal{D}}[\ell(\Psi^{(T)}; P, z)] \le \varepsilon.$$

For distribution-shifted outliers, each test outlier direction is assumed to be a positive linear combination of training outlier directions:

$$v = \sum_{i=1}^{V} \lambda_i v_i^*, \qquad \sum_i \lambda_i \ge L > 0.$$

The learned gate therefore remains small on such outliers, with magnitude controlled by the lower and upper bounds on κ_a'. The condition

$$\alpha < \min(1, p_a l_{\text{tr}}/l_{\text{ts}})$$

and the lower bound on l_{ts} ensure enough clean context examples remain. The same margin calculation then gives the stated test 0-1 error bound.

The mechanism corollaries are the two estimates used inside the proof: linear attention concentrates on the same relevant pattern,

$$\sum_{i\in\mathcal{N}_1} \widetilde{p}_i^\top (W_B^{(T)})^\top W_C^{(T)} \widetilde{p}_{\text{query}} \ge \Theta(1),$$

while the sum over the complement is at most $\Theta((1-p_a)^{-1}\varepsilon)$; the gate suppresses outliers and decays geometrically with distance from the query for clean examples. □

A.7.6 Low-dimensional Mamba ICL

Proof of Oh et al. [31]. The data are Gaussian single-index prompts

$$x \sim \mathcal{N}(0, I_d), \qquad y = g_*(\langle \beta, x \rangle) + \zeta,$$

where β lies on an unknown r-dimensional coordinate subspace. The source fixes admissible maxima $N^*, T^* \le d^{C^*}$ for context length and pretraining tasks. The embedding $\phi(x)$ contains Hermite features up to degree two. The Mamba layer is simplified so that

$$W_B^\top W_C = \mathrm{diag}(\gamma, 0), \qquad w = (0, \dots, 0, \rho^{-1}),$$

and the final Mamba scalar is

$$\mathrm{Mamba}(Z; \gamma)[\widetilde{d}+1, N+1] = \sum_{j=1}^{N} G_{j,N+1}(Z) y_j \phi(x_j)^\top (\gamma \odot \phi(x)).$$

Stage I performs one gradient step on γ. Differentiating the pretraining square loss gives an empirical average of products of Hermite feature vectors, weighted by labels and by the sigmoid gate. The Gaussian-chaos estimate in Appendix A.6.1 replaces this empirical average by its population expectation, so with high probability

$$\gamma^* = 2\eta \mathbb{E}_{\beta \sim \mathrm{Unif}(S_r)}[\psi(\beta, a_0, a_1, a_2) \odot \psi(\beta, b_0, b_1, b_2)] + \text{controlled error}.$$

The sigmoid gate changes the effective Hermite exponent: the relevant coefficient appears at the generative exponent $\mathrm{ge}(g_*)$, not necessarily at the information exponent. Substituting γ^* into the Mamba output and applying concentration over the context examples gives

$$N^{-1}\,\mathrm{Mamba}(Z; \gamma^*)[\widetilde{d}+1, N+1] = P_1 + P_2 \left(\frac{\langle \beta, x \rangle}{r} \right)^{\mathrm{ge}(g_*)} + o\Big(P_2 r^{-3\,\mathrm{ge}(g_*)/2} (\log d)^{-2\deg(g_*)+2} \Big).$$

This is the formal test-time feature-learning statement.

Stage II fixes γ^* and trains the outer MLP. Using polynomial approximation by a two-layer ReLU network, one constructs an outer parameter u' with bounded norm that approximates the link function on the learned scalar feature. The regularized optimization returns u^* with no larger norm and training error at most $\tau + o(1)$.

For test error, define the MLP class with $\|u\| \le U$. Its Rademacher complexity satisfies

$$\mathrm{Rad}_{T_2}(\mathcal{F}_U) = \widetilde{O}(U m^{1/2} T_2^{-1/2}).$$

With $U = \widetilde{O}(r^{3\,\mathrm{ge}(g_*)/2} m^{-1/2})$ and $T_2 = \widetilde{\Omega}(r^{3\,\mathrm{ge}(g_*)})$, the generalization error is $o(1)$. A final comparison between prompts of length N_{pt} and N_{test} uses the concentration of the normalized Mamba output to show their predictions differ by $o(1)$. Therefore

$$\mathcal{R}_{N_{\mathrm{test}}}(\gamma^*, u^*, v^*, a^*) \le \tau + o(1)$$

with probability at least 0.99. □

A.8 Zero-Order-Hold Discretization

Consider the continuous-time LTI system

$$\dot{h}(t) = Ah(t) + Bu(t), \qquad y(t) = Ch(t) + Du(t).$$

Fix a step size $\Delta > 0$ and assume the input is held constant on the interval $((t-1)\Delta, t\Delta]$, with value x_t. By variation of constants,

$$h(t\Delta) = \exp(\Delta A) h((t-1)\Delta) + \int_0^\Delta \exp((\Delta - s)A) B x_t \, ds.$$

After the change of variables $\tau = \Delta - s$, this becomes

$$h_t = \bar{A} h_{t-1} + \bar{B} x_t, \qquad \bar{A} = \exp(\Delta A), \qquad \bar{B} = \int_0^\Delta \exp(\tau A) B \, d\tau.$$

The integral definition is valid even when A is singular. If A is invertible, then

$$\bar{B} = A^{-1}(\exp(\Delta A) - I)B.$$

In selective SSMs the same calculation is applied at each time step with Δ_t and, when present, B_t depending on the current token.

A.9 Recurrent and Convolutional Forms

For the discrete LTI recurrence

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \qquad y_t = Ch_t + Dx_t,$$

assume $h_0 = 0$. Iterating the recurrence gives

$$h_t = \sum_{r=1}^{t} \bar{A}^{t-r}\bar{B}x_r.$$

Indeed, the formula is true for $t = 1$, and the induction step is

$$h_t = \bar{A}\sum_{r=1}^{t-1} \bar{A}^{t-1-r}\bar{B}x_r + \bar{B}x_t = \sum_{r=1}^{t} \bar{A}^{t-r}\bar{B}x_r.$$

Therefore

$$y_t = Dx_t + \sum_{r=1}^{t} C\bar{A}^{t-r}\bar{B}x_r = Dx_t + \sum_{j=0}^{t-1} K_j x_{t-j}, \qquad K_j = C\bar{A}^j\bar{B}.$$

Thus recurrent evaluation and causal convolution are the same LTI operator. With the unilateral z^{-1} convention, the corresponding transfer function is

$$H(z) = D + C(I - z^{-1}\bar{A})^{-1}\bar{B}.$$

Equivalently, after multiplying inside the resolvent by z, one may write the strictly proper part using $C(zI - \bar{A})^{-1}\bar{B}$ up to the chosen normalization convention.

A.10 Transfer Coordinates and State-Free Recovery

Proof. For the coordinate change, set

$$\widehat{A} = SAS^{-1}, \qquad \widehat{B} = SB, \qquad \widehat{C} = CS^{-1}.$$

Then

$$\begin{aligned}\widehat{H}(z) &= h_0 + \widehat{C}(zI - \widehat{A})^{-1}\widehat{B} \\ &= h_0 + CS^{-1}\big(S(zI - A)S^{-1}\big)^{-1}SB \\ &= h_0 + C(zI - A)^{-1}B = H(z).\end{aligned}$$

Thus the rational transfer function is invariant under state-coordinate changes.

For spectral recovery, expand

$$H_\ell(z) = \sum_{j=0}^{\ell-1} h_j z^{-j}.$$

For $m \geq \ell$ and $z \in \mathbb{T}_m$,

$$\begin{aligned}\mathsf{iFFT}_m\big((H_\ell(z))_{z\in\mathbb{T}_m}\big)_t &= \frac{1}{m}\sum_{z\in\mathbb{T}_m} H_\ell(z)z^t \\ &= \frac{1}{m}\sum_{j=0}^{\ell-1} h_j \sum_{z\in\mathbb{T}_m} z^{t-j}.\end{aligned}$$

The roots-of-unity sum is m when $t = j$ and 0 otherwise, so the expression equals h_t for $0 \leq t < \ell$.

For the tail, use the geometric resolvent expansion:

$$\begin{aligned}\sum_{t=\ell+1}^{\infty} CA^{t-1}Bz^{-t} &= CA^\ell z^{-\ell-1}\sum_{r=0}^{\infty}(Az^{-1})^r B \\ &= CA^\ell z^{-\ell-1}(I - Az^{-1})^{-1}B \\ &= CA^\ell z^{-\ell}(zI - A)^{-1}B.\end{aligned}$$

Finally, if

$$H(z) = h_0 + \sum_i \frac{r_i}{z - \lambda_i},$$

then

$$\frac{r_i}{z - \lambda_i} = r_i \sum_{t \geq 1} \lambda_i^{t-1} z^{-t}$$

where the geometric series converges. Inverting the z-transform gives the diagonal impulse response $h_t = \sum_i r_i \lambda_i^{t-1}$ for $t > 0$. □

A.11 Associativity of the Selective Scan

Write a selective recurrence step as an affine map

$$F_{(\bar{A},b)}(h) = \bar{A}h + b.$$

Composition of two such maps is

$$F_{(\bar{A},b)}(F_{(\bar{A}',b')}(h)) = \bar{A}\bar{A}'h + \bar{A}b' + b,$$

so the induced binary operation is

$$(\bar{A}, b) \circ (\bar{A}', b') = (\bar{A}\bar{A}',\ \bar{A}b' + b).$$

This operation is associative because it is composition of functions. Directly,

$$((\bar{A}, b) \circ (\bar{A}', b')) \circ (\bar{A}'', b'') = (\bar{A}\bar{A}'\bar{A}'',\ \bar{A}\bar{A}'b'' + \bar{A}b' + b),$$
$$(\bar{A}, b) \circ ((\bar{A}', b') \circ (\bar{A}'', b'')) = (\bar{A}\bar{A}'\bar{A}'',\ \bar{A}\bar{A}'b'' + \bar{A}b' + b).$$

The prefix product

$$(\bar{A}_t, b_t) \circ \cdots \circ (\bar{A}_1, b_1)$$

therefore gives the affine map from h_0 to h_t. This is the algebraic reason a time-varying recurrence can be parallelized by scan even when it no longer has a fixed convolution kernel.

A.12 Scalar Mamba Gate Calculation

Take the scalar selective SSM with $A = -1$, $B = 1$, and

$$\Delta_t = \text{softplus}(s_t), \qquad s_t = a^\top x_t.$$

Zero-order-hold discretization gives

$$\bar{A}_t = \exp(-\Delta_t), \qquad \bar{B}_t = \int_0^{\Delta_t} \exp(-\tau)\, d\tau = 1 - \exp(-\Delta_t).$$

Since

$$\exp(-\,\text{softplus}(s_t)) = \frac{1}{1 + \exp(s_t)} = 1 - \sigma(s_t),$$

we have

$$\bar{A}_t = 1 - g_t, \qquad \bar{B}_t = g_t, \qquad g_t = \sigma(s_t).$$

The scalar recurrence is therefore

$$h_t = (1 - g_t)h_{t-1} + g_t x_t.$$

This calculation is exact in the scalar setting above. Its interpretation as a gate depends on the special choice $A = -1$, $B = 1$, and softplus parameterization of Δ_t.

A.13 Hidden-Attention Unrolling

For a time-varying recurrence

$$h_i = \bar{A}_i h_{i-1} + \bar{B}_i x_i, \qquad y_i = C_i h_i,$$

with $h_0 = 0$, induction gives

$$h_i = \sum_{j=1}^{i} \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j x_j,$$

where the empty product is the identity. Multiplying by C_i gives

$$y_i = \sum_{j=1}^{i} C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j x_j.$$

Thus the causal coefficient from token j to output i is

$$\tilde{\alpha}_{ij} = C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j, \qquad j \le i.$$

If the transitions are diagonal, $\bar{A}_k = \mathrm{diag}(a_{k,1}, \dots, a_{k,N})$, then the coefficient decomposes as

$$\tilde{\alpha}_{ij} = \sum_{n=1}^{N} C_{i,n} \bar{B}_{j,n} \prod_{k=j+1}^{i} a_{k,n},$$

in the scalar-input, scalar-output case. This is the coordinatewise form of the hidden-attention interpretation.

A.14 Semiseparable Factorization of an SSM Operator

The same unrolling explains the semiseparable structure. Consider the lower triangular sequence operator with blocks

$$M_{ij} = C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j, \qquad i \ge j.$$

Fix a split point s. For rows $i > s$ and columns $j \le s$, factor the transition product at s:

$$M_{ij} = \underbrace{C_i \left(\prod_{k=s+1}^{i} \bar{A}_k \right)}_{U_i} \underbrace{\left(\prod_{k=j+1}^{s} \bar{A}_k \right) \bar{B}_j}_{V_j}.$$

Thus the off-diagonal block crossing the split factors as $U_i V_j$ through an N-dimensional state space. Its rank is therefore at most N, up to the input-output block dimensions. This is the semiseparable rank bound associated with an N-dimensional recurrence.

For an LTI SSM, the same blocks also depend only on the lag $i - j$, giving a Toeplitz lower triangular operator. For a selective SSM, the factors depend on the input through $\bar{A}_k, \bar{B}_j, C_i$, so the operator is generally data-dependent. In SSD, additional scalar-identity structure further reduces the transition product to a scalar mask times query-value factors, which is the structured intersection with masked attention.

A.15 A Stable-Recurrence Norm Estimate

The simplest stability calculation is already informative. Suppose

$$h_t = \bar{A} h_{t-1} + \bar{B} x_t, \qquad y_t = C h_t,$$

with $h_0 = 0$, $\|\bar{A}\| \le \rho < 1$, and $\|x_t\| \le R$. Then

$$\|h_t\| \le \sum_{j=1}^{t} \|\bar{A}\|^{t-j} \|\bar{B}\| \|x_j\| \le \frac{\|\bar{B}\| R}{1 - \rho}.$$

If two input sequences differ only at time j, then the effect on output y_t for $t \geq j$ is bounded by

$$\|\delta y_t\| \leq \|C\| \, \|\bar{A}\|^{t-j} \|\bar{B}\| \, \|\delta x_j\| \leq \|C\| \, \|\bar{B}\| \rho^{t-j} \|\delta x_j\|.$$

This elementary estimate is not a substitute for the paper-specific generalization bounds. It records the common control-theoretic intuition: strict stability gives bounded state norms and exponential decay of old input perturbations.

References

[1] Ameen Ali, Itamar Zimerman, and Lior Wolf. "The Hidden Attention of Mamba Models". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. 2025, pp. 1516–1534.

[2] Marco Bondaschi, Nived Rajaraman, Xiuying Wei, Razvan Pascanu, Caglar Gulcehre, Michael Gastpar, and Ashok Vardhan Makkuva. "From Markov to Laplace: How Mamba In-Context Learns Markov Chains". In: *International Conference on Learning Representations*. 2026.

[3] Yifang Chen, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. "The Computational Limits of State-Space Models and Mamba via the Lens of Circuit Complexity". In: *Conference on Parsimony and Learning*. PMLR. 2025, pp. 739–767.

[4] Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. "Theoretical Foundations of Deep Selective State-Space Models". In: *Advances in Neural Information Processing Systems* 37 (2024).

[5] Edo Cohen-Karlik, Itamar Zimerman, Liane Galanti, Ido Atad, Amir Globerson, and Lior Wolf. "On the Expressivity of Selective State-Space Layers: A Multivariate Polynomial Approach". In: *International Conference on Artificial Intelligence and Statistics*. 2026.

[6] Adam Dankowiakowski and Alessandro Ronca. "Metric Automata Theory: A Unifying Theory of RNNs". In: *Advances in Neural Information Processing Systems* 38 (2025).

[7] Tri Dao and Albert Gu. "Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 10041–10071.

[8] Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. "Hungry Hungry Hippos: Towards Language Modeling with State Space Models". In: *International Conference on Learning Representations*. 2023.

[9] Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. "Unlocking State-Tracking in Linear RNNs Through Negative Eigenvalues". In: *International Conference on Learning Representations*. 2025.

[10] Albert Gu and Tri Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". In: *First Conference on Language Modeling*. 2024.

[11] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. "HiPPO: Recurrent Memory with Optimal Polynomial Projections". In: *Advances in Neural Information Processing Systems* 33 (2020).

[12] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. "On the Parameterization and Initialization of Diagonal State Space Models". In: *Advances in Neural Information Processing Systems* 35 (2022).

[13] Albert Gu, Karan Goel, and Christopher Ré. "Efficiently Modeling Long Sequences with Structured State Spaces". In: *International Conference on Learning Representations*. 2022.

[14] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. "Combining Recurrent, Convolutional, and Continuous-time Models with Linear State-Space Layers". In: *Advances in Neural Information Processing Systems* 34 (2021).

[15] Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. "How to Train Your HiPPO: State Space Models with Generalized Orthogonal Basis Projections". In: *International Conference on Learning Representations*. 2023.

[16] Ankit Gupta, Albert Gu, and Jonathan Berant. "Diagonal State Spaces are as Effective as Structured State Spaces". In: *Advances in Neural Information Processing Systems* 35 (2022).

[17] Ankit Gupta, Harsh Mehta, and Jonathan Berant. "Simplifying and Understanding State Space Models with Diagonal Linear RNNs". In: *arXiv preprint arXiv:2212.00768* (2022).

[18] John T. Halloran, Manbir Gulati, and Paul Roysdon. "Mamba State-Space Models Are Lyapunov-Stable Learners". In: *Transactions on Machine Learning Research* (2025).

[19] Arya Honarpisheh, Mustafa Bozdag, Octavia Camps, and Mario Sznaier. "Generalization Error Analysis for Selective State-Space Models Through the Lens of Attention". In: *Advances in Neural Information Processing Systems* 38 (2025).

[20] Jerry Yao-Chieh Hu, Xiwen Zhang, Ali ElSheikh, Weimin Wu, and Han Liu. "On Structured State-Space Duality". In: *arXiv preprint arXiv:2510.04944* (2025).

[21] Ningyuan Teresa Huang, Miguel Sarabia, Abhinav Moudgil, Pau Rodriguez, Luca Zappella, and Federico Danieli. "Understanding Input Selectivity in Mamba: Impact on Approximation Power, Memorization, and Associative Recall Capacity". In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. PMLR, 2025, pp. 25693–25727.

[22] Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. "Repeat After Me: Transformers are Better than State Space Models at Copying". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 21502–21521.

[23] Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. "Trained Mamba Emulates Online Gradient Descent in In-Context Linear Regression". In: *Advances in Neural Information Processing Systems* 38 (2025).

[24] Arjun Karuvally, Franz Nowak, Anderson T. Keller, Carmen Amo Alonso, Terrence J. Sejnowski, and Hava T. Siegelmann. "Bridging Expressivity and Scalability with Adaptive Unitary SSMs". In: *Advances in Neural Information Processing Systems* 38 (2025).

[25] Behnoush Khavari, Mehran Shakerinava, Jayesh Khullar, Jerry Huang, François Rivest, Siamak Ravanbakhsh, and Sarath Chandar. "Parity Requires Unified Input Dependence and Negative Eigenvalues in SSMs". In: *arXiv preprint arXiv:2508.07395* (2025).

[26] Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. "Mamba-3: Improved Sequence Modeling using State Space Principles". In: *International Conference on Learning Representations*. 2026.

[27] Hongkang Li, Songtao Lu, Xiaodong Cui, Pin-Yu Chen, and Meng Wang. "Can Mamba Learn In Context with Outliers? A Theoretical Generalization Analysis". In: *arXiv preprint arXiv:2510.00399* (2025).

[28] Fusheng Liu and Qianxiao Li. "From Generalization Analysis to Optimization Designs for State Space Models". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 31383–31405.

[29] William Merrill, Jackson Petty, and Ashish Sabharwal. "The Illusion of State in State-Space Models". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 35492–35506.

[30] Vinoth Nandakumar, Qiang Qu, Peng Mi, and Tongliang Liu. "State space models can express n-gram languages". In: *Transactions on Machine Learning Research* (2025).

[31] Junsoo Oh, Wei Huang, and Taiji Suzuki. "Mamba Can Learn Low-Dimensional Targets In-Context via Test-Time Feature Learning". In: *arXiv preprint arXiv:2510.12026* (2025).

[32] Rom N. Parnichkun et al. "State-Free Inference of State-Space Models: The Transfer Function Approach". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 39834–39860.

[33] Dániel Rácz, Mihály Petreczky, and Bálint Daróczy. "Length independent generalization bounds for deep SSM architectures via Rademacher contraction and stability constraints". In: *Transactions on Machine Learning Research* (2025).

[34] Yuval Ran-Milo, Eden Lumbroso, Edo Cohen-Karlik, Raja Giryes, Amir Globerson, and Nadav Cohen. "Provable Benefits of Complex Parameterizations for Structured State Space Models". In: *Advances in Neural Information Processing Systems* 37 (2024).

[35] Yash Sarrof, Yana Veitsman, and Michael Hahn. "The Expressive Capacity of State Space Models: A Formal Language Perspective". In: *Advances in Neural Information Processing Systems* 37 (2024).

[36] Mehran Shakerinava, Behnoush Khavari, Siamak Ravanbakhsh, and Sarath Chandar. "The Expressive Limits of Diagonal SSMs for State-Tracking". In: *International Conference on Learning Representations*. 2026.

[37] Mugunthan Shandirasegaran, Hongkang Li, Songyang Zhang, Meng Wang, and Shuai Zhang. "A Theoretical Analysis of Mamba's Training Dynamics: Filtering Relevant Features for Generalization in State Space Models". In: *International Conference on Learning Representations*. 2026.

[38] Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. "Simplified State Space Layers for Sequence Modeling". In: *International Conference on Learning Representations*. 2023.

[39] Neeraj Mohan Sushma, Yudou Tian, Harshvardhan Mestha, Nicolo Colombo, David Kappel, and Anand Subramoney. "State-space models can learn in-context by gradient descent". In: *arXiv preprint arXiv:2410.11687* (2024).

[40] Aleksandar Terzić, Michael Hersche, Giacomo Camposampiero, Thomas Hofmann, Abu Sebastian, and Abbas Rahimi. "On the Expressiveness and Length Generalization of Selective State-Space Models on Regular Languages". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39.19 (2025), pp. 20876–20884.

[41] Aleksandar Terzić, Nicolas Menet, Michael Hersche, Thomas Hofmann, and Abbas Rahimi. "Structured Sparse Transition Matrices to Enable State Tracking in State-Space Models". In: *Advances in Neural Information Processing Systems* 38 (2025).

[42] Shida Wang and Qianxiao Li. "StableSSM: Alleviating the Curse of Memory in State-space Models through Stable Reparameterization". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 50766–50793.